Universidade de São Paulo
Faculdade de medicina de Ribeirão Preto

A BAYESIAN APPROACH FOR LEFT-CENSORED DATA BASED ON MIXTURE AND SEMI-CONTINUOUS MODELS USING TOBIT STRUCTURE

Danielle Peralta

Ribeirão Preto
2022

Danielle Peralta

A Bayesian approach for left-censored data based on mixture and semi-continuous models using Tobit structure

Tese de doutorado apresentada ao Programa de Saúde Pública da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para a obtenção do título de Doutor em Ciências.

"Versão corrigida. A versão original encontra-se disponível tanto na Biblioteca da Unidade que aloja o Programa, quanto na Biblioteca Digital de Teses e Dissertações da USP (BDTD)"

Universidade de São Paulo - FMRP/USP

Supervisor: Jorge Alberto Achcar

Ribeirão Preto

2022

Catalog record electronically elaborated by the author, using the LaTeX template of USPSC class from University of São Paulo.

Peralta, Danielle

A Bayesian approach for left-censored data based on mixture and semi-continuous models using Tobit structure/ Danielle Peralta. – Ribeirão Preto, 2022.

137p.

Dissertation (Ph.D.) – Medical School of Ribeirão Preto. Universidade de São Paulo - FMRP/USP . Supervisor: Achcar, Jorge Alberto.

1. Bayesian approach. 2. Tobit models. 3. Left-censoring. 4. Medical studies. 5. Public health. 6. Regression models. 7. Environmental, medical and astronomy data. I. Achcar, Jorge Alberto. II. University of São Paulo. III. Medical School of Ribeirão Preto. IV. A Bayesian approach for left-censored data based on mixture and semi-continuous models using Tobit structure.

Ficha catalográfica elaborada eletronicamente pelo autor, utilizando o modelo LaTeX da classe USPSC da Universidade de São Paulo.

Peralta, Danielle
Uma abordagem Bayesiana para dados censurados à esquerda baseada em modelos de mistura e semi-contínuos usando a estrutura Tobit. / Danielle Peralta. – Ribeirão Preto, 2022.
137p.

Tese (Doutorado) – Faculdade de Medicina de Ribeirão Preto. Universidade de São Paulo - FMRP/USP. Orientador: Achcar, Jorge Alberto.

1. Análise Bayesiana. 2. Modelos Tobit. 3. Censura à esquerda. 4. Estudos médicos. 5. Saúde pública. 6. Modelos regressão. 7. Dados ambientais, médicos, e astronômicos. I. Achcar, Jorge Alberto. II. Universidade de São Paulo. III. Faculdade de Medicina de Ribeirão Preto. IV. Uma abordagem Bayesiana para modelos de mistura e semi-contínuos dados com censuras à esquerda usando a estrutura de Tobit.

Peralta, D. **A Bayesian approach for left-censored data based on mixture and semi-continuous models using Tobit structure**. 137p. Dissertation (Ph.D.) – Medical School of Ribeirão Preto, Universidade de São Paulo - FMRP/USP , 2022.

Examining Board

Jorge Alberto Achcar
(Supervisor/President)

Jair Licio Ferreira Santos
(1^{st} examiner)

Wesley Bertoli da Silva
(2^{nd} examiner)

Marcos V. de Oliveira Peres
(3^{rd} examiner)

Ribeirão Preto, 2022.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Código de Financiamento 001

Ao meu filho, Benício.

Acknowledgements

Agradeço ao orientador, o Professor Doutor Jorge Alberto Achcar pela orientação desta tese, pela inspiração, pelo apoio e pela sua disposição ao longo desses quatro anos. E principalmente por entender minhas limitações, muito obrigada.

Agradeço ao amigo, o Professor Doutor Ricardo Puziol de Oliveira, pelo apoio e parceria nos artigos.

Agradeço aos membros da banca, ao professor Doutor Jair Licio Ferreira Santos, Doutor Wesley Bertoli da Silva e Doutor Marcos V. de Oliveira Peres por aceitarem o convite de compor a banca examinadora.

Agradeço ao Departamento de Medicina Social da Faculdade de Medicina de Ribeirão em que o trabalho foi desenvolvido e por todos seus docentes, pela oportunidade que me deram de fazer parte desta pós-graduação.

Agradeço à Secretaria do Departamento de Medicina Social, em especial a Paula, pela ajuda prestada frente as dúvidas e burocracias exigidas neste período.

Enfim, quero demonstrar o meu agradecimento, a todos aqueles que, de um modo, tornaram possível a realização da presente tese. Muito obrigada!

Abstract

A Bayesian approach for left-censored data based on mixture and semi-continuous models using Tobit structure

The main objective of this thesis is to introduce a left-censored data analysis using the tobit model for univariate and multivariate data. The tobit model can be used as an alternative to the least squares regression model when the assumption of linearity is not satisfied. The tobit model is able to fit the data adequately by formulating a regression model for which the response is pre-fixed to a limit value. In this thesis we present five chapters, each considering a manuscript submitted for publication and with different approaches and applications. The estimation of the model parameters is performed using Bayesian inference methods. The summaries a *posterior*i of interest are obtained using existing MCMC (Monte Carlo on Markov Chains) simulation methods, as Gibs and Metropolis-Hasting. In the first paper (Chapter 2) we present the tobit-Weibull mixture model to analyze environmental data under the left censoring scheme. The considered dataset is related to ammonia nitrogen concentrations in rivers. In the second paper (Chapter 3), the bivariate tobit-Weibull model under a hierarchical Bayesian analysis is presented considering a dataset in stellar astronomy where a fragility or latent variable is considered to capture the possible correlation between the bivariate responses for the same sample unit; applications of the univariate and bivariate tobit-Weibull model are also presented in Chapter 4, considering two medical datasets (cancer survival data and vaccine data). The tobit-Weibull model in the presence of some covariates with linear and quadratic effects, under the left censoring scheme, is presented in Chapter 5 considering a dataset concerning total daily precipitation collected at a weather station located in the city of São Paulo, Brazil. In Chapter 6 we present a generalized form of the tobit-Weibull model in the presence of covariates and excess zeros; the application was performed using data concerning total daily precipitation. Chapter 7 concludes this thesis with general conclusions showing the usefulness of the proposed model fot analyzing left-censored data or with an excess of zero-valued observations.

Keywords: *Tobit model, left censored data, Bayesian analysis, MCMC methods, Weibull distribution, data analysis.*

Resumo

Uma abordagem Bayesiana para dados censurados à esquerda baseada em modelos de mistura e semi-contínuos usando a estrutura Tobit

O principal objetivo desta tese é introduzir uma análise de dados censurada à esquerda usando o modelo tobit para dados univariados e multivariados. O modelo tobit pode ser usado como uma alternativa ao modelo de regressão de mínimos quadrados quando a suposição de linearidade não é satisfeita. O modelo proposto é capaz de se ajustar adequadamente aos dados, formulando um modelo de regressão para o qual a resposta é pré-fixada a um valor limite. Nesta tese, apresentamos cinco capítulos, cada um considerando um manuscrito submetido para publicação e com diferentes abordagens e aplicações. A estimativa dos parâmetros do modelo é feita usando métodos de inferência Bayesianos. Os resumos a *posteriori* de interesse são obtidos usando os métodos de simulação existentes MCMC (Monte Carlo on Markov Chains), como Gibs e Metropolis-Hasting. No primeiro trabalho (Capítulo 2) apresentamos o modelo de mistura tobit-Weibull para analisar os dados ambientais. O conjunto de dados considerado está relacionado às concentrações de nitrogênio amônia em rios. No segundo trabalho (Capítulo 3), é apresentado o modelo tobit-Weibull bivariado sob uma análise Bayesiana hierárquica considerando um conjunto de dados em astronomia estelar onde uma variável de fragilidade ou latente é considerada para capturar a possível correlação entre as respostas bivariadas para a mesma unidade amostral. Aplicações do modelo univariado e bivariado tobit-Weibull também são apresentadas no Capítulo 4, considerando dois conjuntos de dados médicos (dados de sobrevivência ao câncer e dados de vacinas). O modelo tobit-Weibull na presença de alguns covariáveis com efeitos lineares e quadráticos é apresentado no Capítulo 5, considerando um conjunto de dados referentes à precipitação total diária coletada em uma estação meteorológica localizada na cidade de São Paulo, Brasil. No Capítulo 6 apresentamos uma forma generalizada do modelo tobit-Weibull na presença de covariáveis e excesso de zeros; a aplicação foi realizada utilizando dados referentes à precipitação total diária. O Capítulo 7 conclui esta tese com conclusões gerais mostrando a utilidade do modelo proposto para análise de dados censurados à esquerda ou com um excesso de observações com valor nulo.

Palavras-chave: Modelo de Tobit, dados com censuras à esquerda, análise Bayesiana, métodos MCMC, distribuição de Weibull , análise de dados.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Background

In survival analysis, the main goal is to analyze the time until the occurence of an event of interest. The data for this kind of analysis is called time-to-event data. A classic example of time-to-event data is the time until the death of a patient, as the term suggests, but it can also be any well defined characteristic. For example, in medical research, events of interest can be: the onset of Alzheimer's disease, the recurrence time of a cancer, the time of exposed individuals becoming infected, the time for a patient to be free of a disease. In reliability analysis, the event of interest can be linked to the time until the failure of a particular component of a system, or be related to the breakdown and repair of machines. In other areas of study the possible events are: the time of probation of criminals (criminology); time of service, time of marriage until divorce (sociology); time of hospitalization, time until a company's bankruptcy (administration) among many other applications (see, for example, Cox, 1972, Maller and Zhou, 1996, Klein and Moeschberger, 1997, De Angelis et al., 1999, Lee and Wang, 2003, Fleming and Lin, 2000, Frees, 2009, Cox et al., 2007, Giolo and Colosimo, 2006, Lagakos and Williams, 1978, Bewick et al., 2004, Struthers and Farewell, 1989, Leung et al., 1997, Lindsey and Ryan, 1998, Guo, 1993, Lynn, 2001, Romeu, 2004, Rausand and Hoyland, 2004, Ibrahim et al., 2005, Sreeja and Sankaran, 2008, Hougaard, 2012, Crowder, 2012, Eryilmaz and Tank, 2012).

An important characteristic of survival data occurs when some individuals in the study do not experience the event of interest at the end of the study or withdraw during the follow-up period of analysis. For example, some patients may still be alive or free of disease at the end of the study period. The exact survival time of these individuals is unknown. The data of these individuals are called censored observations or censored

times occurring for several reasons such as: the patient may drop-out of the study before the observation of the event of interest, or die of other causes different of the goal of the study. Without the presence of censoring, classical classical statistical techniques, such as regression analysis, could be used in the analysis of the data [Giolo and Colosimo, 2006].

A survival time is censored if all that is known is that it started or ended within some given time interval, and thus the total cycle length (from entry time to transition) is not known exactly. Censoring can occur in a number of ways, among which are:

- Type I censoring: occur in studies that when terminated after a predetermined period of time record, without their termination, some individuals who have not yet presented the event of interest. As an example, a clinical study in which the event of interest is the death of a subject after being diagnosed with a certain malignant tumor; if the individual is alive at the end of the study, you have censoring on the right;
- Type II censoring: a sample of n units are tested, for which the experiment begins at a fixed time zero, $t = 0$, and ends when a fixed number of units, $r(r < n)$, have failed. Failure times are only observed for r units, i.e., units that fail after unit r has failed are not observed. The total number of censored units is fixed, while the experimental time is random. Experiments involving Type II censoring are often used to test equipment life;
- Type III or random censoring: typically occurs in time-to-event medical studies. An individual who withdraws from the study before the event of interest occurs has a random censoring value. For example, the subject may change address, no longer want to participate in the study, or when the participate in the study, or when the subject dies for a reason other than the one other than the one studied.

1.1.1 Left-Censored Data

To define this scheme mathematically, let Y be a random variable denoting the lifetime of an unit or patient such that the lifetime data is given by $T = \max(C, Y)$ where C is a censored time and Y is a complete observation. Thus, we could define a indicator variable as

$$\delta = \begin{cases} 1, & \text{if T is a complete observation } (Y > C) \\ 0, & \text{if T is a left censored observation } (Y \leq C). \end{cases} \tag{1.1}$$

Examples of left-censored data:

- Left-censoring is very common in environmental analysis. For example, in water quality studies, the censorship could occurs when the level of a chemical trace in a sample is less than the "limit of quantification" (LOQ) or "limit of detection" (LOD) of the analytical instruments. The physical meanings of LOQ and LOD differ associated to the analytical technology applied. Such observations are usually reported as "less than detectable", meaning that a measurement was made, but its low level prevented the reporting of a quantitative value [Akritas et al., 1994]. The literature on the analysis of censored environmental data is largely driven by the issues raised by water quality analysis. Water quality problems are widespread where negative health impacts are associated even with low levels of concentrations of certain chemicals. Among several articles already published using frequentist and Bayesian methods to model viral concentrations we can cite, for example, Petterson et al. [2015], Pouillot et al. [2015], Vergara et al. [2016], Atwood et al. [1991].

- In the medical field, a relevant topic of study is the determination of antibody concentration by quantitative assays. This topic is relevant because there is always a concentration value (threshold) where below this value an accurate measurement cannot be obtained, regardless of the technique used. When dealing with data from an assay where left-censoring is present, the lower limit of detection (LD) can be used to replace the unobserved value as a censored observation. In this regard, one can cite the studies of Moulton and Halsey, 1995, Lynn, 2001, Guo, 1993, Balakrishnan, 1989, Balakrishnan and Varadan, 1991, Arellano-Valle et al., 2012, Canales et al., 2018, Achcar et al., 2018, Mitra and Kundu, 2008, Jacqmin-Gadda et al., 2000.

1.1.2 Generalized forms of the Weibull distribution

A very popular distribution widely used in reliability studies is the Weibull distribution [Weibull, 1951], mainly due to the flexibility of its hazard function and the facility to estimate its parameters. In data analysis considering positive asymmetric data, new classes of parametric distributions based on extensions of the Weibull distribution have been introduced in the literature. As special cases, we have the exponentiated Weibull (EW) (Mudholkar and Srivastava, 1993, Pal et al., 2006), the generalized modified Weibull [Carrasco et al., 2008] and the log-beta Weibull distributions [Ortega et al., 2013].

Some generalized forms of the Weibull distribution can be seen in a review paper

introduced by [Pham and Lai, 2007] among which we can mention Gurvich et al. [1997] that introduced a class of distributions generalizing the traditional two parameters Weibull distribution; Nadarajah and Kotz [2006] proposed another generalization of the Weibull distribution that contains the model proposed by Xie et al. [2002], with the model proposed by Chen [2000] as a particular case; Muralidharan and Lathika [2006] considered a lifetime situation with early failures showing that such situation can be modeled by mixing a Weibull distribution with a singular distribution, thus resulting in a generalized Weibull model; Nikulin and Haghighi [2006] proposed a generalized power Weibull distribution with three parameters. Also a modified form of the Weibull distribution was introduced by Lai et al. [2003].

1.1.2.1 The generalized modified Weibull (GMW) distribution

A generalized modified Weibull (GMW) [Carrasco et al., 2008] distribution with four parameters is defined by a probability density function given by,

$$f(t) = \frac{\alpha\beta t^{\gamma-1}(\gamma+\lambda t)\exp\left(\lambda t - \alpha t^{\gamma}e^{\lambda t}\right)}{\left\{1-\exp\left(-\alpha t^{\gamma}e^{\lambda t}\right)\right\}^{(1-\beta)}} \tag{1.2}$$

where $t > 0$, α, β, γ and λ are positive parameters and the survival function $S(t) = P(T > t)$ is given by,

$$S(t) = 1 - \left\{1-\exp\left(-\alpha t^{\gamma}e^{\lambda t}\right)\right\}^{\beta} \tag{1.3}$$

The GMW distribution with four parameters contains as special cases many usual lifetime distributions as the Weibull distribution when $\lambda = 0$ and $\beta = 1$; the exponential distribution when $\lambda = 0$, $\beta = 1$ and $\gamma = 1$; the Rayleigh distribution when $\lambda = 0$, $\beta = 1$ and $\gamma = 2$; the extreme value distribution when $\beta = 1$ and $\gamma = 0$; the Exponentiated Weibull distribution (EW) when $\lambda = 0$; the Exponentiated exponential distribution (EE) when $\lambda = 0$ and $\gamma = 1$; the Generalized Rayleigh distribution (GR) when $\lambda = 0$ and $\gamma = 2$ and the Modified Weibull distribution (MW) when $\beta = 1$.

1.1.2.2 The exponentiated Weibull distribution

A exponentiated Weibull (EW) [Mudholkar and Srivastava, 1993] distribution with three parameters is obtained from (6.4) assuming $\lambda = 0$, with probability density function given by,

$$f(t) = \frac{\alpha\beta\gamma t^{\gamma-1}\exp\left(-\alpha t^{\gamma}\right)}{\left\{1-exp\left(-\alpha t^{\gamma}\right)\right\}^{1-\beta}} \tag{1.4}$$

where $t > 0$, α, β and γ are positive parameters and the survival function $S(t) = P(T > t)$ is given by,

$$S(t) = 1 - \{1 - \exp(-\alpha t^{\gamma})\}^{\beta} \tag{1.5}$$

1.1.2.3 The Weibull distribution

The Weibull distribution [Weibull, 1939], widely known for its simplicity and flexibility in accommodating different forms of risk function, is perhaps the most widely used distribution model for life time analysis. For a random variable T with Weibull distribution, the probability density function is given by,

$$f(t) = \frac{\alpha}{\beta^{\alpha}} t^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^{\alpha}\right\}, t \geq 0 \tag{1.6}$$

where α is the shape parameter and β the scale parameter, both positive. For this distribution, the survival function $S(t) = P(T > t)$ and the hazard function $h(t)$ are given respectively by,

$$S(t) = \exp\left\{-\left(\frac{t}{\beta}\right)^{\alpha}\right\} \tag{1.7}$$

where $t > 0$ and $\alpha > 0$, $\beta > 0$. The mean of the Weibull distribution with density (1.6) is given by $E(T) = \beta\Gamma(1 + 1/\alpha)$. In this case, one may have increasing risks (failure rates) if $\alpha > 1$; decreasing if $\alpha < 1$ and constant if $\alpha = 1$, that is, we have great flexibilty of it for the data.

Another parameterization of the Weibull distribution is obtained from (6.4) assuming $\lambda = 0$ and $\beta = 1$ (a Weibull distribution with two parameters), with probability density function given by,

$$f(t) = \alpha\gamma t^{\gamma-1} \exp(-\alpha t^{\gamma}) \tag{1.8}$$

where $t > 0$, α and γ are positive parameters and the survival function $S(t) = P(T > t)$ is given by,

$$S(t) = \exp(-\alpha t^{\gamma}) \tag{1.9}$$

Bayesian inferences for the three distributions introduced in this section are presented by many authors in the literature (see, for example, Achcar et al. [1985, 1999] and Martinez et al. [2013]).

1.1.3 Tobit Models

The Tobit model was proposed by Tobin [1958] for a limited (censored) dependent variable (or response). Tobin was motivated to develop his model by a case study where he needed to analyze the relationship between household expenditure on a durable good with household incomes. The common regression approach with ordinary least squares could not be used in this situation because there were many cases where the expenditure was zero, which destroyed the assumption of linearity. To solve the problem, Tobin proposed a model that could fit the data appropriately formulating a regression model whose response was censored to a prefixed limiting value (see Amemiya [1984]).

1.1.3.1 Tobit Models for Left-Censored Data

Let Y be an independent random variable and $Y = (Y_1, \ldots, Y_m, Y_{m+1}, \ldots, Y_n)^\top$ be a sample of size n. Suppose that this sample includes m censored observations and $n - m$ uncensored observations. Thus, such censoring scheme can be visualized under a regression setting with a censored response Y^*, which is a (unobserved) latent variable. Hence, the m censored data (unobserved) correspond to the values of Y^* less than or equal to a threshold point y_0, so that all of these data take the value y_0 (censoring to the left). The other $n - m$ data (observed) are related to values of Y^* greater than y_0, which can be described by a linear regression structure of the type $\mathbf{x}_i^\top \boldsymbol{\beta}$. This modeling approach may be formulated by the normal Tobit model with censored response to the left as

$$Y_i = \begin{cases} y_0, & Y_i^* \leq y_0 \qquad i = 1, \ldots, m \\ Y_i^* = \mathbf{x}_i^\top \beta + \varepsilon_i, & Y_i^* > y_0 \qquad i = m+1, \ldots, n \end{cases} \tag{1.10}$$

where $\varepsilon_i \sim N(0, \sigma^2)$ is the model error term, β is a vector of regression coefficients corresponding to unknown parameters to be estimated, and $\mathbf{x}_i$ is a vector containing the covariate values. Observe that y_0 given in (2.1) is a prefixed limiting value that makes the response of the regression model to be limited (or censored), as mentioned by Tobin [1958].

Tobit models rely on the normality assumption. Proposals of Tobit models that relax this assumption are extremely important, since it is of common knowledge that most of the data available in the real world are often well modeled by non-normal distributions. A number of authors have noticed that the asymmetry of data for censored responses and their kurtosis usually are different from the expected for a normal distribution, so that more flexible Tobit models are needed. The interested reader is referred to Barros et al., 2010, 2016, Arellano-Valle et al., 2012, Chib, 1992, Martínez-Flórez et al., 2013, Leiva et al.,

2007, Villegas et al., 2011, Moulton and Halsey, 1995, Amemiya, 1984, Thorarinsdottir and Gneiting, 2010, Desousa, 2016 for some works related to non-normal Tobit models.

1.2 Motivation

Mixture models provide a way to model time to failure in various situations where a single parametric probability density is inadequate to correctly describe the heterogeneity of the data. In the medical field, mixture models are applied, for example, in studies of diseases with multiple stages of development, where the time to failure at each stage is modeled by a different parametric family (see e.g. Boag [1949]) or when a proportion of patients recovering after treatment can be defined and estimated (see, for example, [De Angelis et al., 1999]. Mixture models allow one to build probabilistic models in a wide variety of phenomena in various fields of knowledge, e.g., engineering, economics [Mosler, 2003] and hydrology, among others (see, Titterington et al. [1985], McLachlan and Peel [2000]).

Interest in these models has increased in survival analysis, especially in studies related to cancer treatment. In general, survival analysis models assume that all individuals in the population studied are susceptible to the event of interest and will eventually experience the event if follow-up is long enough. These data can arise from clinical trials in which, even after prolonged follow-up, no further events of interest are observed. Some people in the population may be considered cured (or not susceptible). In recent years, there has been growing interest in modeling survival data for long-term survivors.

Most work has focused on the mixture model with cure fraction, where it is assumed that the study population is a mixture of susceptible individuals who experience the event of interest and individuals who will never experience that event. These individuals are not at risk with respect to the event of interest and are considered immune, non-susceptible, or cured [Maller and Zhou, 1996]. In light of this, estimating the proportion of cured has also become a highly relevant objective. The works presented by Boag [1949] and Berkson and Gage [1952], introduced the standard mixture model, which formed the basis of what came to be called the long-term survival model or the survival model with fraction of cure. Different approaches, parametric and nonparametric, have been considered as a model for the proportion of immune (see, for example, Haybittle, 1965, Farewell, 1982, Tsodikov, 1998, Price and Manatunga, 2001, Cancho and Bolfarine, 2001, Tsodikov et al., 2003, Yu et al., 2004, Yin and Ibrahim, 2005, Lambert et al., 2006, Lu, 2010, Othus et al., 2012, Achcar et al., 2012, 2013, Fernandes, 2014).

1.3 Goals

The main objective of this thesis is in the definition, characterization and comparisons of mixture models in survival analysis for left-censored data, extending the number of possible marginal distributions for its different components. The specific goals are:

- To introduce Tobit-Weibull models as well the main mathematical properties and inferences under a Bayesian approach using MCMC methods to get the *posterior* summaries of interest.
- To provide different analyses for data with left censoring scheme using the proposed model and comparisons with other existing model approaches.
- To use statistical softwares (R and OpenBUGS), for reproducible research where the computer codes are available for other researchers working with this class of models.

1.4 Organization of chapters

The thesis is organized as follows: In Chapter 2, it is presented our first study about Tobit-Weibull model based on a mixture approach to analyze environment data under left-censoring scheme. The dataset considered in this study is related to ammonia nitrogen concentrations (in mg/L) in rivers located in the Washington State in the period ranging from 2011 to 2016. Also, simulation studies were carried out to illustrate the performance of the parameter estimators of the proposed model. As expected, the results showed that the Tobit-Weibull model could be useful to describe the behavior of the ammonia nitrogen concentrations as well to predict the probabilities of those concentrations.

In Chapter 3, it is presented the bivariate Tobit-Weibull model under a hierarchical Bayesian analysis of a stellar astronomy dataset. A frailty or latent variable is considered to capture the possible correlation between the bivariate responses for the same sampling unit. The *posterior* summaries of interest are obtained using existing (MCMC) methods. A comparison of the two models using the different likelihood approaches (Weibull or Weibull-Tobit likelihoods) also is discussed in the application.

In Chapter 4, it is presented the use of the proposed models in other fields of study as, for example, medical data analysis. To accomplish our goal, we considered two datasets: cancer survival data and vacine data. In this way, using a left-censored medical data related

to differentiated thyroid cancer, we fitted the Tobit-Weibull models in two ways: assuming the data as univariate and bivariate. For instance, the dataset consists in 91 patients and was used by López et al. [2014] in a descriptive study to evaluate the relationship between the initial thyroglobulin levels and the presence of recurrence of cancer one year after receiving treatment. For the second dataset, the goal of the study was to investigate that the higher titer vaccines could effectively immunize infants as young as 6 months of age. Neutralization antibody assays were performed in children at 12 months of age, the dataset was used by Moulton and Halsey [1995].

In Chapter 5, the Tobit-Weibull model in the presence of some covariates with linear and quadratic effects, under left censoring scheme, is presented. The data set considered is related to total daily rainfall collected at a climate station located in the city of São Paulo, Brazil, in the period from 2007 to 2021. Under a Bayesian approach using Markov Chain Monte Carlo methods to obtain the *posterior* summaries of interest, we also simultaneously used a logistic regression model for the occurrence (or not) of daily rainfall over the follow-up time period. Other climate variables such as daily mean atmospheric pressure, daily mean temperature, and daily mean humidity are also analyzed over the follow-up time period.

In Chapter 6, are presented Tobit-generalized Weibull models in the presence of covariates and excess zeros. A special application of the proposed models is considered with daily rainfall data obtained from a climate station in the city of São Paulo, Brazil over the period 2007 to 2021.

Finally, Chapter 7 end this thesis with general conclusions on all studies presented here from where reinforce the fact that the search of appropriate statistical model could be extremely difficult depending on the censoring structure of the lifetime data. However, the proposed methodology could be very useful in the medical data analysis in presence of left-censored scheme.

Chapter 2

Environmental data under a left-censoring mechanism: An application to river ammonia nitrogen concentrations using Tobit-Weibull model

2.1 Introduction

Survival data, *lifetime data, failure time data,* or *time-to-event data* are terms used to describe data that measure the time to the occurrence of some event which arises in a number of applied fields. In medical research, the events of interest might be, for example, the response time to a treatment or length of stay in the hospital. In reliability analysis, the event of interest can be related to the time until a machine shuts down. Survival models can also be used in other applications that do not involve "time to event", for example, in water quality studies to check if the water is suitable for human consumption; in medical studies to check the levels of a certain substance increased/decreased in the blood after a treatment. In these two cases, survival models can be used because of the asymmetry of the distribution, and also because they have an important feature of survival analysis, which is censoring.

Censoring in these cases is common because the response variable is measured by analytical instruments. For example, in water quality studies, censoring can occur when the level of a chemical trait in a sample is less than the "detection limit" of the analytical instrument used. Such values are reported as "less than detectable", i.e. there is a measurement, but its low level has prevented the reporting of a quantitative value. Such

observations are treated as left censoring.

This Chapter it is presented our first study about Tobit-Weibull model based on a mixture approach to analyze environment data under left-censoring scheme. It is organized as follows: In Sections 2.2 the Tobit model and the Tobit-Weibull model are presented and some associated inference method. Section 2.3 reports the results from a simulation study done to evaluate the performance of the proposed estimation procedure. Section 2.4 presents an application with a real water river quality dataset related to the concentrations (mg/L) of ammonia nitrogen (NH3-N) in the rivers located in the Washington state, USA, in a specified period of time, between the years of 2011 and 2016, where the response of interest is the amount of ammonia (NH3-N). Finally, Section 2.5 concludes the paper with some comments and remarks.

2.2 Materials and Methods

Tobit Models for Left-Censored Data

Let $Y = (Y_1, \ldots, Y_m, Y_{m+1}, \ldots, Y_n)^\top$ be a sample of size n, that is, independent random variables but not necessarily independent identically distributed. Assume that this sample includes m left-censored observations and $n-m$ observed (complete or uncensored) data. Thus, such censoring scheme can be visualized under a regression setting with a censored response Y^*, which is a (unobserved) latent variable. Hence, the m censored data (unobserved) correspond to the values of Y^* less than or equal to a threshold point y_0 (censoring to the left), so that all of these data take the value y_0. The other $n-m$ data (observed) are related to values of Y^* greater than y_0, which can be described by a linear regression structure of the type $\mathbf{x}_i^\top \boldsymbol{\beta}$. This modeling approach may be formulated by the normal Tobit model with censored response to the left as

$$Y_i = \begin{cases} y_0, & Y_i^* \leq y_0 \qquad i = 1, \ldots, m \\ Y_i^* = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, & Y_i^* > y_0 \qquad i = m+1, \ldots, n \end{cases} \tag{2.1}$$

where $\varepsilon_i \sim N(0, \sigma^2)$ is the model error term, β is a vector of regression coefficients corresponding to unknown parameters to be estimated, and $\mathbf{x}_i$ is a vector containing the covariate values. Observe that y_0 given in (2.1) is a prefixed limiting value that makes the response of the regression model to be limited (or censored), as mentioned by Tobin [1958].

Tobit models rely on the normality assumption. Proposals of Tobit models that relax this assumption are extremely important, since it is of common knowledge that most

of the data available in the real world are often well modeled by non-normal distributions. A number of authors have noticed that the asymmetry of data for censored responses and their kurtosis usually are different from the expected for a normal distribution, so that more flexible Tobit models are needed. The interested reader is referred to Barros et al., 2010, 2016, Arellano-Valle et al., 2012, Chib, 1992, Martínez-Flórez et al., 2013, Leiva et al., 2007, Villegas et al., 2011, Moulton and Halsey, 1995, Amemiya, 1984, Thorarinsdottir and Gneiting, 2010, Desousa, 2016 for some works related to non-normal Tobit models.

Remark 1. *T*he Tobit and probit models are similar in many ways. Each one of them have the same structural model, just different measurement models that is, how the Y^* is translated into the observed y is different in each model. In the Tobit model, we know the value of Y^* when $Y^* > y_0$, while in the probit model we only know if $Y^* > y_0$. Since there is more information in the Tobit model, the estimates of the regression parameters β's should be more efficient. The interested reader is referred to [Long et al., 1997] for details about logit, probit and Tobit models.

Tobit-Weibull Model

From the censoring indicator defined by (1.1), we have, $\delta = 1$ if T is a complete observation $(Y > C)$ and $\delta = 0$ if T is a left-censored observation $(Y \leq C)$. If we have a complete observation, that is, $(Y > C)$, let us assume a truncated Weibull distribution with probability density function given by,

$$f(t \mid T > C) = \frac{f_0(t)}{S_0(t)} \tag{2.2}$$

where

$$f_0(t) = \frac{\alpha}{\beta^\alpha} t^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\} \quad \text{and} \quad S_0(t) = P(T > t) = \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\}$$

where $t > 0$ and $\alpha > 0$, $\beta > 0$.

Remark 2. *T*he baseline survival function given by

$$S_0(C) = P(T > C) = \exp\left\{-\left(\frac{C}{\beta}\right)^\alpha\right\}$$

where C is a known constant.

Let us assume the mixture model, given by the probability density function

$$f(t) = p + (1-p)\frac{f_0(t)}{S_0(C)} \tag{2.3}$$

where p is the mixing parameter,

$$f_0(t) = \frac{\alpha}{\beta^\alpha} t^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\}$$

and

$$S_0(C) = \exp\left\{-\left(\frac{C}{\beta}\right)^\alpha\right\}.$$

For the proposed model, we could observe that:

- If $T \leq C$, then

$$S(t) = P(T > t) = p + (1-p)\int_C^\infty \frac{f_0(u)du}{S_0(C)} = p + (1-p)\frac{S_0(C)}{S_0(C)} = p + (1-p) = 1$$

- If $T > C$, then

$$S(t) = P(T > t) = 1 - P(T \leq C) = 1 - \{P(0 < T < C) + P(C < T < t)\}$$
$$= 1 - \left\{p + (1-p)\int_C^t \frac{f_0(u)du}{S_0(C)}\right\}$$

where

$$\int_C^t f_0(u)du = P(T > C) - P(T > t) = S_0(C) - S_0(t).$$

Thus,

$$S(t) = 1 - \left\{p + \frac{(1-p)}{S_0(C)}[S_0(C) - S_0(t)]\right\} = 1 - \left\{p + (1-p) - [(1-p)\frac{S_0(t)}{S_0(C)}]\right\}.$$

That is,

$$S(t) = (1-p)\frac{S_0(t)}{S_0(C)}.$$

In summary, the survival function of the proposed Tobit-Weibull model is given by

- If $T \leq C$, $S(t) = 1$
- If $T > C$, $S(t) = (1-p)\frac{S_0(t)}{S_0(C)}$

where $S_0(t) = \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\}$ and $S_0(C) = \exp\left\{-\left(\frac{C}{\beta}\right)^\alpha\right\}$.

Inference Method for Tobit-Weibull Model

The likelihood function for the parameters p, α and β based on one observation is given from (2.3) by

$$L(p,\alpha,\beta) = p + (1-p)\frac{f_0(t)}{S_0(C)} \tag{2.4}$$

where

$$f_0(t) = \frac{\alpha}{\beta^\alpha} t^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\}$$

and

$$S_0(C) = \exp\left\{-\left(\frac{C}{\beta}\right)^\alpha\right\}.$$

Now, with the censoring information (1.1), let us define a binary variable $\delta = 1$ if T is a complete observation ($T > C$) and $\delta = 0$ if T is a left-censored observation ($T \leq C$) with conditional probabilities given by

$$P(\delta = 0 \mid p,\alpha,\beta,t) = \frac{p}{p + (1-p)\dfrac{f_0(t)}{S_0(C)}} \tag{2.5}$$

which could be assumed as a Bernoulli trial. In this way, the likelihood function for n observations, $L(p,\alpha,\beta;t)$, is simply given by

$$L(p,\alpha,\beta;t) = \prod_{i=1}^{n} p^{(1-\delta_i)} \left[(1-p)\frac{f_0(t_i)}{S_0(C)}\right]^{\delta_i}. \tag{2.6}$$

Now, assuming the truncated Weibull distribution, the likelihood and the log-likelihood functions for the proposed Tobit-Weibull model p, α and β are given respectively (from (2.6)) by

$$L(p,\alpha,\beta;t) = \prod_{i=1}^{n} p^{(1-\delta_i)} \left[(1-p)\frac{\frac{\alpha}{\beta^\alpha} t_i^{\alpha-1} \exp\left\{-\left(\frac{t_i}{\beta}\right)^\alpha\right\}}{\exp\left\{-\left(\frac{C}{\beta}\right)^\alpha\right\}}\right]^{\delta_i}$$

and

$$\begin{aligned}\ell(p,\alpha,\beta;t) &= \sum_{i=1}^{n}\Bigg\{(1-\delta_i)\log(p) + \delta_i\Bigg\{\log(1-p) + \log(\alpha) + \\ &+ (\alpha-1)\log(t_i) - \alpha\log(\beta) - \left(\frac{t_i}{\beta}\right)^\alpha + \left(\frac{C}{\beta}\right)^\alpha\Bigg\}\Bigg\}.\end{aligned}$$

For a Bayesian approch of the proposed models, we use MCMC simulation methods, based on both Gibbs and Metropolis–Hastings sampling, this approach is better known

as Metropolis-within-Gibbs algorithms, to get the *posterior* summaries of interest (see, for example, Chib and Greenberg, 1995, Gelfand and Smith, 1990, Gelman et al., 1995, Geman and Geman, 1984, Gilks et al., 1995). From where it as assumed Gamma(a, b) *prior* distributions for the parameters α and β with a and b known hyperparameters and a Beta(e, f) distribution for p with e and f known hyperparameters. Moreover, in presence of a vector of covariates $\mathbf{x} = (x_1, x_2, \ldots, x_p)^\top$ let us assume a regression model for the scale parameter β given by

$$\boldsymbol{\beta_i} = \exp(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \ldots + \gamma_p x_p)$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_p)^\top$ is the regression parameter vector associated to covariate vector $\mathbf{x} = (x_0, x_1, x_2, \ldots, x_p)^\top$ and a logistic model for the parameter p, given by

$$logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \varphi_0 + \varphi_1 x_1 + \varphi_2 x_2 + \ldots + \varphi_p x_p. \tag{2.7}$$

In this case, it as assumed a Gamma(a, b) prior distributions for the parameter α, Normal(c, d^2) prior distributions for the regression parameters $\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_p$ with c and d known hyperparameters and Normal(e, f^2) *prior* distributions for the regression parameters $\varphi_0, \varphi_1, \varphi_2, \ldots, \varphi_p$ with e and f known hyperparameters.

2.3 A Simulation Study

In Section 2.2, it was considered a Bayesian approach as inference method to get the estimators of the parameters of the proposed models assuming approximately non-informative *priors*. Alternatively, it was used maximum likelihood estimation (MLE) methods to get the estimators for the parameters of interest. In this section, it is presented the results of a simulation study to evaluate the performance of a MLE estimation procedure to get the estimators of the proposed models assuming different scenarios and different sample sizes.

In this way, we present a Monte Carlo simulation study with 1000 replications to evaluate the performance of the MLE of the Weibull truncated model parameters, using the R Software [R Development Core Team, 2009]. The sample sizes considered are $n = 50, 100, 150, 200, 250, 300$, with parameters $\alpha = 0.5, 1.0$, $\beta = 0.5, 1.0, 4.0$ and censoring proportions equal to $p = 0.1, 0.3, 0.5, 0.7$. We compute the empirical bias and mean squared errors (MSE) in order to present the performance evaluation. Tables 1, 2, 3 and 4 present the obtained results for the indicated sample sizes, parameters values and

censoring proportions. Note that in all tables, the empirical bias and MSE decrease when n increases, as expected.

Table 1 – Estimated BIAS/MSE for the MLE estimators of the parameters α, β and p considering $\alpha = 0.5$, $\beta = 0.5$, $c = 1.0$ and proportion of censoring given by ρ

	$\rho = 0.1$						$\rho = 0.3$					
	Bias			MSE			Bias			MSE		
n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$
50	0.3704	0.0759	0.0022	0.7115	0.0599	0.0018	0.4334	0.0958	-0.0031	0.8888	0.0830	0.0046
100	0.2096	0.0387	-0.0002	0.3494	0.0281	0.0009	0.2482	0.0496	-0.0024	0.4295	0.0359	0.0022
150	0.1443	0.0245	0.0001	0.2177	0.0172	0.0006	0.1670	0.0311	-0.0004	0.2678	0.0208	0.0015
200	0.1201	0.0210	-0.0001	0.1616	0.0123	0.0004	0.1252	0.0212	0.0001	0.2009	0.0155	0.0011
250	0.0927	0.0161	-0.0003	0.1239	0.0093	0.0004	0.1113	0.0199	0.0002	0.1692	0.0129	0.0009
300	0.0733	0.0123	-0.0007	0.1028	0.0077	0.0003	0.0912	0.0157	0.0002	0.1380	0.0104	0.0007

	$\rho = 0.5$						$\rho = 0.7$					
	Bias			MSE			Bias			MSE		
n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$
50	0.5282	0.1423	-0.0016	1.1664	0.1409	0.0048	0.8508	0.2704	0.0011	2.1061	0.3181	0.0042
100	0.3227	0.0668	-0.0015	0.6192	0.0554	0.0024	0.4882	0.1207	0.0000	1.0112	0.1002	0.0021
150	0.2187	0.0421	-0.0002	0.3885	0.0334	0.0016	0.3487	0.0764	0.0001	0.6586	0.0579	0.0014
200	0.1869	0.0370	0.0003	0.2970	0.0244	0.0012	0.2754	0.0587	-0.0010	0.4831	0.0428	0.0010
250	0.1481	0.0277	0.0004	0.2376	0.0191	0.0010	0.2443	0.0489	-0.0004	0.3982	0.0332	0.0009
300	0.1246	0.0233	-0.0001	0.1888	0.0149	0.0008	0.2202	0.0436	-0.0002	0.3279	0.0267	0.0007

Table 2 – Estimated BIAS/MSE for the MLE estimators of the parameters α, β and p considering $\alpha = 1.0$, $\beta = 0.5$, $c = 1.0$ and proportion of censoring given by ρ

	$\rho = 0.1$						$\rho = 0.3$					
	Bias			MSE			Bias			MSE		
n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$
50	0.4539	0.0666	0.0011	1.2760	0.0393	0.0017	0.4843	0.0716	0.0011	1.6041	0.0551	0.0044
100	0.2548	0.0339	-0.0003	0.6705	0.0182	0.0008	0.3072	0.0403	0.0007	0.8626	0.0246	0.0023
150	0.1802	0.0227	-0.0004	0.4467	0.0112	0.0006	0.2031	0.0260	-0.0003	0.5756	0.0155	0.0015
200	0.1243	0.0152	-0.0003	0.3366	0.0082	0.0004	0.1541	0.0209	-0.0004	0.4411	0.0118	0.0011
250	0.1024	0.0131	-0.0006	0.2659	0.0065	0.0004	0.1208	0.0161	0.0001	0.3336	0.0089	0.0009
300	0.0822	0.0104	-0.0005	0.2343	0.0057	0.0003	0.1130	0.0150	0.0009	0.2805	0.0073	0.0008

	$\rho = 0.5$						$\rho = 0.7$					
	Bias			MSE			Bias			MSE		
n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$
50	0.6355	0.1016	-0.0004	2.2790	0.0851	0.0051	1.0053	0.1958	-0.0025	4.0277	0.1995	0.0042
100	0.3652	0.0518	-0.0020	1.0917	0.0337	0.0025	0.6073	0.0949	-0.0005	2.0067	0.0700	0.0022
150	0.2614	0.0362	-0.0019	0.7328	0.0208	0.0017	0.4371	0.0655	0.0000	1.3122	0.0434	0.0014
200	0.2049	0.0274	-0.0008	0.5617	0.0152	0.0013	0.3572	0.0508	-0.0001	1.0346	0.0311	0.0011
250	0.1759	0.0240	-0.0004	0.4666	0.0127	0.0011	0.2890	0.0403	-0.0005	0.7860	0.0228	0.0009
300	0.1535	0.0214	0.0004	0.3963	0.0106	0.0009	0.2320	0.0315	0.0001	0.6091	0.0176	0.0007

Table 3 – Estimated BIAS/MSE for the MLE estimators of the parameters α, β and p considering $\alpha = 1.0$, $\beta = 1.0$, $c = 1.0$ and proportion of censoring given by ρ

	$\rho = 0.1$						$\rho = 0.2$					
	Bias			MSE			Bias			MSE		
n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$
50	0.0968	0.1320	0.0019	0.2438	0.1624	0.0017	0.1187	0.1653	-0.0010	0.3076	0.2233	0.0041
100	0.0311	0.0516	0.0014	0.1343	0.0646	0.0009	0.0522	0.0786	-0.0014	0.1790	0.0945	0.0021
150	0.0223	0.0353	0.0011	0.0982	0.0422	0.0006	0.0391	0.0567	-0.0008	0.1282	0.0627	0.0015
200	0.0190	0.0279	0.0008	0.0740	0.0313	0.0005	0.0251	0.0420	-0.0001	0.0994	0.0455	0.0011
250	0.0116	0.0211	0.0006	0.0625	0.0260	0.0004	0.0223	0.0337	0.0000	0.0848	0.0363	0.0009
300	0.0069	0.0154	0.0002	0.0523	0.0214	0.0003	0.0240	0.0316	0.0008	0.0750	0.0314	0.0008

	$\rho = 0.3$						$\rho = 0.5$					
	Bias			MSE			Bias			MSE		
n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$
50	0.1465	0.2107	0.0046	0.3886	0.3487	0.0050	0.2319	0.4011	0.0006	0.5251	0.7799	0.0038
100	0.0646	0.0979	0.0024	0.2402	0.1439	0.0025	0.0996	0.1668	-0.0008	0.3286	0.2632	0.0021
150	0.0386	0.0582	0.0014	0.1595	0.0786	0.0016	0.0660	0.1152	0.0012	0.2661	0.1767	0.0014
200	0.0338	0.0480	0.0013	0.1258	0.0595	0.0012	0.0624	0.0946	0.0021	0.2099	0.1217	0.0011
250	0.0255	0.0385	0.0012	0.1077	0.0490	0.0010	0.0446	0.0676	0.0011	0.1776	0.0891	0.0008
300	0.0155	0.0289	0.0010	0.0896	0.0393	0.0008	0.0384	0.0591	0.0006	0.1570	0.0767	0.0007

Table 4 – Estimated BIAS/MSE for α, β and p considering $\alpha = 1.0$, $\beta = 4.0$, $c = 1.0$ and proportion of censoring given by ρ

	$\rho = 0.1$						$\rho = 0.3$					
	Bias			MSE			Bias			MSE		
n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$
50	-0.0137	0.4888	-0.0003	0.0284	2.6509	0.0018	-0.0249	0.5725	0.0019	0.0407	3.5365	0.0041
100	-0.0069	0.2316	-0.0002	0.0129	1.1009	0.0009	-0.0177	0.2515	0.0002	0.0200	1.5571	0.0020
150	-0.0064	0.1425	-0.0005	0.0083	0.7108	0.0006	-0.0140	0.1441	-0.0016	0.0124	0.9702	0.0014
200	-0.0035	0.1244	-0.0010	0.0060	0.5484	0.0004	-0.0103	0.1093	-0.0008	0.0090	0.7095	0.0010
250	-0.0015	0.1113	-0.0008	0.0046	0.4420	0.0003	-0.0098	0.0726	-0.0011	0.0071	0.5671	0.0008
300	-0.0028	0.0762	-0.0011	0.0038	0.3578	0.0003	-0.0085	0.0478	-0.0007	0.0053	0.4454	0.0007

	$\rho = 0.5$						$\rho = 0.7$					
	Bias			MSE			Bias			MSE		
n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}$
50	-0.0244	0.8678	-0.0019	0.0523	5.6754	0.0051	-0.0413	1.2999	0.0016	0.0829	11.6471	0.0043
100	-0.0143	0.4650	0.0002	0.0274	2.4039	0.0025	-0.0369	0.5362	0.0017	0.0491	4.0309	0.0020
150	-0.0123	0.2881	0.0005	0.0176	1.5071	0.0017	-0.0243	0.3435	0.0010	0.0289	2.3142	0.0013
200	-0.0120	0.1982	0.0006	0.0132	1.1103	0.0012	-0.0184	0.2467	-0.0003	0.0213	1.5676	0.0010
250	-0.0081	0.1662	0.0000	0.0100	0.8534	0.0009	-0.0149	0.1927	-0.0003	0.0161	1.2318	0.0008
300	-0.0048	0.1491	0.0004	0.0078	0.7052	0.0008	-0.0143	0.1587	-0.0001	0.0140	1.0829	0.0007

2.4 Data Application: River Ammonia Nitrogen

The Washington State Department of Ecology monitoring team collects water samples from more than 85 long-term rivers and stream stations with 67 stations classified as long-term, 8 stations classified as sentinel and 12 stations classified as basin. The freshwater monitoring team collects 24-hour data for dissolved oxygen, temperature, pH, and conductivity in many rivers and streams statewide. They also collect monthly data on bacteria, pH, phosphorus, and more. This data displays long-term trends in stream health and contributes to watershed studies and water quality improvement plans.

For our analysis, it is considered the concentrations (mg/L) of ammonia nitrogen (NH3-N) in the rivers located in the Washington state, USA between the years of 2011 and 2016. Also, some risk factors are assumed as, for example, pH, oxygen concentration, nitrite and nitrate (NO_2-NO_3) concentration, pressure, temperature, turbidity, among others. The dataset was obtained from the website https://ecology.wa.gov/. To summarize the descriptive results, Figure 1 presents some boxplots related to the ammonia nitrogen concentration and the water parameters grouped by eco region. From the boxplots in Figure 1, it can be seen that for the eco region 3, Columbia Basin, there is higher variability in the covariates, nitrogen dioxide and nitrate concentrations, phosphorus (sol reactive) concentration, dissolved oxygen concentration, water pH, water barometric pressure, water temperature and total persulfate nitrogen when compared to the other eco regions.

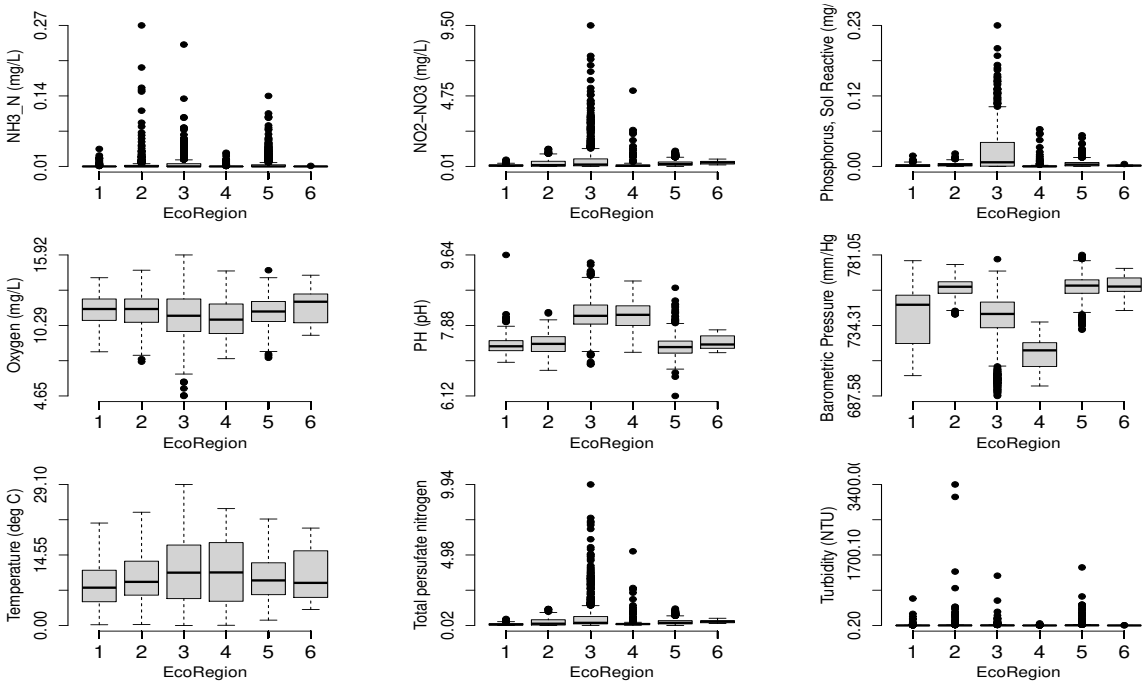

Figure 1 – Boxplots for the ammonia nitrogen concentration and the water parameters grouped by Eco Region.

In the sample considered in this study, there are $n = 3141$ observations where 2090 observations are left-censored data in the value 0.010 and 1051 observations are complete observations (values larger than 0.010). The regions of the United States are assumed in the regression models considering "dummy" variables denoting the Eco Regions (Cascades = 1 and 0 for other eco region; Coast R. = 1 and 0 for other regions; Columbia B. = 1 and 0 for other regions; Northerm R. = 1 and 0 for other regions; Puget L. = 1 and 0 for other regions; the Willamet V. eco region is the reference). From the dataset it is observed $2090/3141 = 0,6654$ or $66,54\%$ of left-censored data.

Use of the Tobit-Weibull model

For the data analysis, we assumed the proposed model and, as *prior* distributions, an approximately non-informative Gamma(0.01, 0.01)*prior* distribution for α and β an informative $Beta(7, 3)$ *prior* distribution for the parameter p based on the elicitation of an informative *prior* distribution given by Carlin and Louis [2010]. Table 5 presents the *posterior* summaries of interest based on the final Gibbs sample of size 3,000 choosen among every 100 simulated sample (300,000 simulations) and considering a burn-in sample of size 11,000 to eliminate the initial effects of the parameters in the iterative procedure. Table 5 also presents the maximum likelihood estimators (MLE) and the 95% confidence

intervals for the parameters of both assumed models, from where it is observed similar results as obtained under a Bayesian approach.

Table 5 – *Posterior* summaries and MLE of interest for the Tobit-Weibull model.

Model	Parameter	*posterior* Mean (SD)	95% Credible Interval	MLE (SD)	95% Confidence Interval
Tobit-Weibull	α	0.3957 (0.0485)	(0.3023; 0.4939)	0.3687 (0.0525)	(0.2658, 0.4716)
	β	0.0006 (0.0004)	(0.0001; 0.0016)	0.0003 (0.0003)	(-0.0003, 0.0009)
	p	0.6772 (0.0084)	(0.6605; 0.6937)	0.6572 (0.0074)	(0.6428, 0.6716)

Based on the Bayesian estimates for p, α and β, we can write the estimated survival function of the proposed model for the estimation of the survival curve (a common feature when there is the presence of censoring mechanisms). In this way, assuming the Bayesian estimates, we have that,

- $S(t) = 1$ if $t \leq 0.01$,
- $S(t) = (1 - 0.6772) \exp\{-(t^{0.3957} - c^{0.3957})/0.0006\}$ if $t > 0.01$

Now, in order to identify which water covariate affects the response ammonia nitrogen concentration, a regression approach is considered in the presence of the following covariates:

- $NO_2 - NO_3$: nitrogen dioxide and nitrate concentrations;
- OP-DIS: phosphorus (sol reactive) concentration;
- Oxygen(O_2): dissolved oxygen concentration
- pH: water pH;
- Press: water barometric pressure;
- Temp: water temperature (in ^oC)
- TPN: total persulfate nitrogen;
- TURB: turbidity;
- EcoRegion: Cascades (Eco1), Coast Range (Eco2), Columbia Basin (Eco3), Northerm Rockies (Eco4), Puget Lowland (Eco5), Willamet Valley (Ref).

The regression structure for the proposed model, in this case, is based on a linear model for β and a logistic model for p, and is given by,

$$\begin{aligned}\beta_i &= \exp(\gamma_0 + \gamma_1(NO_2 - NO_3)_i + \gamma_2(OP - DIS)_i + \gamma_3(Oxygen)_i + \gamma_4(pH)_i + \gamma_5(Press)_i \\ &+ \gamma_6(Temp)_i + \gamma_7(TPN)_i + \gamma_8(Turb)_i + \gamma_9(Eco1)_i + \gamma_{10}(Eco2)_i + \gamma_{11}(Eco3)_i \\ &+ \gamma_{12}(Eco4)_i + \gamma_{13}(Eco5)_i)\end{aligned}$$

and,

$$logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \varphi_0 + \varphi_1(NO_2 - NO_3)_i + \varphi_2(OP - DIS)_i + \ldots + \varphi_{13}(Eco5)_i$$

In this case, an approximately non-informative $Gamma(0.01, 0.01)$ *prior* distribution for the parameter α, an approximately non-informative $N(0, 100)$ *prior* distributions for the regression parameters $\gamma_0, \gamma_1, \gamma_2, \ldots \gamma_{13}$ and an approximately $N(0, 100)$ *prior* distributions for the regression parameters $\varphi_0, \varphi_1, \varphi_2, \ldots, \varphi_{13}$ were assumed for better computational stability. Table 6, presents the *posterior* summaries of interest based on the final Gibbs sample of size 2,000 chosen amongst every 100 simulated sample (200,000 simulations) and considering a burn-in sample of size 11,000 to eliminate the initial effects of the parameters in the iterative procedure as well the MLE estimator.

Table 6 – *Posterior* summaries and MLE of interest for the Tobit-Weibull model with regression structure

Parameter		Mean (SD)	95% Credible Interval	MLE (SD)	95% Confidence Interval
Shape	α	0.7308 (0.0662)	(0.5955; 0.8579)	0.6944 (0.0635)	(0.5699; 0.8188)
Scale	γ_0 (Intercept)	-1.1747 (1.0456)	(-2.8703; 0.7495)	1.4488 (20.2119)	(-38.1706; 41.0682)
	γ_1 ($NO_2 - NO_3$)	-2.0140 (0.4417)	(-2.7485; -1.0314)	-3.5631 (0.5430)	(-4.6274; -2.4988)
	γ_2 (OP-DIS)	2.3763 (0.8741)	(0.6540; 4.0682)	0.2613 (0.0686)	(0.1268; 0.3957)
	γ_3 (Oxygen)	-0.1584 (0.0341)	(-0.2274; -0.0908)	-0.1986 (0.0568)	(-0.3100; -0.0872)
	γ_4 (pH)	0.0225 (0.1207)	(-0.2393; 0.2175)	-0.0019 (0.1788)	(-0.3524; 0.3486)
	γ_5 (Press)	-0.0034 (0.0014)	(-0.0061; -0.0010)	-0.7554 (3.0859)	(-6.8044; 5.2936)
	γ_6 (Temp)	-0.0358 (0.0085)	(-0.0520; -0.0192)	-0.0468 (0.0150)	(-0.0763; -0.0173)
	γ_7 (TPN)	2.0792 (0.4199)	(1.1693; 2.7756)	3.4159 (0.5340)	(2.3692; 4.4626)
	γ_8 (Turb)	0.0023 (0.0004)	(0.0016; 0.0032)	0.1782 (0.0323)	(0.1149; 0.2415)
	γ_9 (Eco1)	-0.2233 (0.4309)	(-1.0184; 0.6207)	1.6020 (0.6293)	(0.3685; 2.8355)
	γ_{10} (Eco2)	0.0983 (0.4188)	(-0.7093; 0.8459)	1.7089 (0.5947)	(0.5432; 2.8747)
	γ_{11} (Eco3)	-0.1177 (0.4145)	(-0.8950; 0.6741)	1.1566 (0.6002)	(-0.0199; 2.3331)
	γ_{12} (Eco4)	-0.5037 (0.4429)	(-1.2704; 0.3664)	1.2918 (0.6591)	(-0.0001; 2.5838)
	γ_{13} (Eco5)	0.2307 (0.4010)	(-0.5488; 0.9676)	1.5901 (0.5839)	(0.4456; 2.7345)
Mixing	φ_0 (Intercept)	0.8275 (0.0491)	(0.7369; 0.8916)	0.8483 (21.8122)	(-41.9079; 43.6046)
	φ_1 ($NO_2 - NO_3$)	0.1784 (0.0187)	(0.1266; 0.2029)	11.2456 (1.1454)	(9.0003; 13.4909)
	φ_2 (OP-DIS)	-1.4587 (0.1436)	(-1.7466; -1.1714)	-0.9873 (0.0737)	(-1.1317; -0.8428)
	φ_3 (Oxygen)	0.0113 (0.0009)	(0.0095; 0.0134)	0.2307 (0.0687)	(0.0960; 0.3654)
	φ_4 (pH)	-0.0001 (0.0026)	(-0.0052; 0.0049)	0.6544 (0.1854)	(0.2910; 1.0179)
	φ_5 (Press)	-0.0004 (0.0001)	(-0.0005; -0.0002)	-1.6395 (3.3034)	(-8.1149; 4.8360)
	φ_6 (Temp)	0.0012 (0.0003)	(0.0005; 0.0018)	-0.0203 (0.0188)	(-0.0571; 0.0165)
	φ_7 (TPN)	-0.1718 (0.0170)	(-0.1958; -0.1246)	-11.3945 (1.1103)	(-13.5709; -9.2181)
	φ_8 (Turb)	-0.0002 (0.0000)	(-0.0002; -0.0002)	-0.1889 (0.0359)	(-0.2591; -0.1186)
	φ_9 (Eco1)	0.0087 (0.0071)	(-0.0009; 0.0251)	0.1822 (0.4592)	(-0.7179; 1.0823)
	φ_{10} (Eco2)	-0.0088 (0.0080)	(-0.0229; 0.0082)	-0.6544 (0.4266)	(-1.4905; 0.1818)
	φ_{11} (Eco3)	-0.0022 (0.0081)	(-0.0136; 0.0175)	0.0071 (0.4575)	(-0.8897; 0.9039)
	φ_{11} (Eco4)	-0.0116 (0.0096)	(-0.0280; 0.0100)	-0.6179 (0.5177)	(-1.6328; 0.3969)
	φ_{13} (Eco5)	-0.0187 (0.0073)	(-0.0301; -0.0005)	-0.6092 (0.4193)	(-1.4311; 0.2126)

Based on the results assuming the Bayesian estimates in Table 6, we could observe that the ammonia nitrogen is affected by $NO_2 - NO_3$ concentration (γ_1), phosphorus concentration (γ_2), dissolved oxygen concentration (γ_3), water barometric pressure (γ_5), water temperature (γ_6), total persulfate nitrogen (γ_7), turbidity (γ_8) assuming the linear structure; and $NO_2 - NO_3$ concentration (φ_1), phosphorus concentration (φ_2), dissolved oxygen concentration (φ_3), water barometric pressure (φ_5), water temperature (φ_6), total persulfate nitrogen (φ_7), turbidity (φ_8), Puget Lowland eco region (φ_{13}) assuming logistic structure.

Assuming the linear structure which is our interest, it is observe negative relationships for dissolved oxygen and water temperatures, which implies that when dissolved oxygen concentration and water temperature decreases, the ammonia nitrogen increases. This result is in accord to Fatimah et al. [2017]. Same behavior occurs to the nitrogen dioxide and nitrate concentrations, phosphorus concentration and water barometric pressure, implying that the extent to which these covariates decrease, the ammonia nitrogen increases. Now, for total persulfate nitrogen and turbidity, there is a positive relationship

that implies that when these covariates increase, the ammonia nitrogen also increases.

The results, assuming the logistic structure is contrary to the linear structure assuming the previous variables related, that is, if there is a negative relationship, then by the logistic structure this relationship becomes positive and vice versa. Both results are important since its depends on the parameter structure of the investigation. If our goal is the scale parameter, then, we assume the results for linear structure; if the parameter of interest is the mixing parameter, then we assume the results for logistic structure.

2.5 Concluding remarks

The main goal of this paper was to propose a new Tobit-Weibull model to identify the main risk factors that affects the ammonia nitrogen concentration for Washington State rivers. For that, it has been considered a regression structure based on linear and logistic models which the main advantage of the this model is the dynamic of the, iteration process and computational stability, to describe the behavior of ammonia nitrogen concentrations. Our approach was based on adopting the month's sequential label from which the ammonia nitrogen concentrations were taken as response.

Moreover, the inclusion of the risk factors provided more accurate model fits; whose underlying results may offer suggestions on how the concentrations for ammonia nitrogen are affected in the considered period and its relationship as highlighted for the significant factors found in the analysis: $NO_2 - NO_3$ concentration, phosphorus concentration, dissolved oxygen concentration, water barometric pressure, water temperature, total persulfate nitrogen, turbidity, Puget Lowland eco region. Nevertheless, the present methodology can also be applied to data from other rivers to provide a comprehensive understanding of the risk factors that affect the ammonia nitrogen concentration, which may alert authorities to keep restrictive strategies to control the advance of this kind of water pollution.

Chapter 3

Bayesian analysis for bivariate Weibull distribution under left-censoring scheme

3.1 Introduction

Many parametric regression models were introduced in the literature to analyse lifetime data in presence of censored data (see for example, [Lawless, 1982]). A very popular semi-parametric regression model extensively used in survival data analysis was introduced by Cox [1972] assuming proportional hazards (see also, Cox and Oakes, 1984, Collett, 2003, Kalbfleisch and Prentice, 2002, Klein and Moeschberger, 1997, Lee and Wang, 2003). In all these models, independent observations are usually assumed, that is, the sample units are not related to each other.

In many applications, especially in medical survival analysis studies, it is possible to have dependent bivariate responses (two or more measurements in the same unit). To capture the correlation between two or more survival times, we could consider the introduction of "frailties" or latent variables (Clayton and Cuzick, 1985, Oakes, 1986, 1989, dos Santos and Achcar, 2011, McGilchrist and Aisbett, 1991, Shih, 1992). Random effects models are largely used to model heterogeneity as the frailty model introduced by Vaupel [1986] used in multivariate survival analysis.

Other possibility in the statistical analysis of bivariate lifetime data is to assume existing parametric probability bivariate lifetime distributions as bivariate exponential, bivariate Weibull, bivariate Lindley or bivariate log-normal distributions (see for example, Gumbel, 1960, Arnold and Strauss, 1988, Block and Basu, 1974, Hougaard, 1986, Marshall and Olkin, 1967, 1985, Downton, 1970, Hawkes, 1972, de Oliveira et al., 2018, Oliveira

et al., 2019, de Oliveira et al., 2021). Other possibility is to use bivariate distributions derived from copula functions (Nelsen, 2007, Trivedi and Zimmer, 2007, Sklar, 1959).

As an example, and motivation for this study, we consider a stellar astronomy bivariate dataset `https://www.iiap.res.in/astrostat/School08/datasets/censor.html` in presence of left censored observations introduced by Santos et al. [2004] (see dataset in Appendix 7 at the end of the manuscript). In this example, the authors seek differences in the properties of stars that do and do not host extrasolar planetary systems where a previously identified sample of objects (stars, galaxies, quasars, X-ray sources, etc.) are observed at some new wavelength or for some new property. This dataset is related to the birth and death of stars where many questions still exist, despite the scientists now understand over 90% of a star's life [Collins, 1989, Chiosi, 1998].

Some of the target objects are detected and the value of the new property is measured, while others are not detected. These are assigned as an upper limit to the value of the property based on the uncertainty of the unsuccessful measurement, that is, we have the presence of left-censored data. The probability to find a planet is a steeply rising function of the star's metal content, but it is unclear whether this arises from the metallicity at birth or from later accretion of planetary bodies. The study introduced by Santos et al. [2004] focuses on two responses associated to the same star: the abundances of the light elements beryllium (Be) and lithium (Li) that are thought to be depleted by internal stellar burning, so that excess of Be and Li should be present only in the planet accretion scenario of metal enrichment. In this way, we have the presence of left censored bivariate data associated to each star.

The main goal of this study is to introduce a hierarchical Bayesian analysis for bivariate Weibull data considering usual Weibull likelihood and Tobit likelihood model approaches based on Weibull distributions for the marginal distributions in presence of left-censoring mechanism. The paper is organized as follows: Section 3.2 presents the proposed Weibull model approaches for bivariate data assuming data with left-censoring mechanism and covariates and inference methods for the parameters of the model. Section 3.3 presents an application of the proposed methodology considering a stellar astronomy data under a hierarchical Bayesian approach. Finally, Section 3.4 closes the paper with some concluding remarks and directions for future research.

3.2 Materials and Methods

Weibull likelihood function considering bivariate data in presence of left-censored data and covariates

Let us assume Weibull distributions [Weibull, 1951] for the univariate responses of interest. The Weibull distribution, widely known for its simplicity and flexibility in accommodating different forms of hazard function, is the most widely used distribution model for lifetime analysis. The Weibull distribution for a random variable T has probability density function given by,

$$f(t) = \frac{\alpha}{\beta^{\alpha}} t^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^{\alpha}\right\}, t \geq 0 \tag{3.1}$$

where α is the shape parameter and β the scale parameter, both positive. Let us denote the Weibull distribution with density (3.1) as Wei(α, β). For this distribution, the survival function $S(t) = P(T > t)$ and the hazard function h(t) are given respectively by,

$$S(t) = \exp\left\{-\left(\frac{t}{\beta}\right)^{\alpha}\right\} \quad \text{and} \quad h(t) = \frac{\alpha}{\beta^{\alpha}} t^{\alpha-1} \tag{3.2}$$

where $t > 0$ and $\alpha > 0$, $\beta > 0$. The mean of the Weibull distribution with density (3.1) is given by $E(T) = \beta\Gamma(1 + 1/\alpha)$ where $\Gamma(\cdot)$ denotes the Gamma function. In this case, one may have increasing risks (failure rates) if $\alpha > 1$; decreasing if $\alpha < 1$ and constant if $\alpha = 1$, that is, we have great flexibilty of fit for the data.

In the analysis of bivariate data (T_1, T_2) in presence of a covariate vector $\mathbf{x} = (x_1, x_2, ..., x_p)^{\top}$ affecting both dependent random variables assuming Weibull distributions Wei(α_1, β_1) and Wei(α_2, β_2), respectively, we consider the use of hierarchical Bayesian methods. In this way, we assume regression models for the scale parameters β_j in the Weibull density (3.1), given by,

$$\beta_{ji} = \exp(\gamma_{j0} + \gamma_{j1} x_{1i} + \gamma_{j2} x_{2i} + \ldots + \gamma_{jp} x_{pi} + w_i) \tag{3.3}$$

where $\boldsymbol{\gamma_j} = (\gamma_{j0}, \gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jp})^{\top}$ is the regression parameter vector associated to the covariate vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^{\top}$, $j = 1, 2; i = 1, 2, \ldots, n$ (sample size); w_i is a random factor which captures extra-Weibull variability and dependence structure between both dependent variables (T_1, T_2). The random factors or latent variables (not-observed) $W_i, i = 1, \ldots, n$, are assumed to be independent random variables with a Normal$(0, \sigma^2)$ distribution.

Assuming a left-censored mechanism, the lifetime data is given by $T_j = \max(C_j, Y_j)$ where C_j is a censored time and Y_j is a complete observation, $j = 1, 2$. Define a censorship indicator variable given by $\delta_j = 1$ if T_j is a complete observation ($Y_j > C_j$) and $\delta_j = 0$ if T_j is a left censored observation ($Y_j \leq C_j$). In this way, the likelihood function based only in one bivariate observation (t_1, t_2) is given by, $F_1(t_1)^{\delta_1 - 1} f_1(t_1)^{\delta_1} F_2(t_2)^{\delta_2 - 1} f_2(t_2)^{\delta_2}$ where $F_j(t_j) = P(T_j \leq t_j) = 1 - S_j(t_j)$ and $f_j(t_j)$ is the probability density function, $j = 1, 2$.

Thus, assuming Weibull distributions Wei(α_1, β_1) and Wei(α_2, β_2) with density (3.1) for the random variables T_1 and T_2 and the regression models (3.3) for the scale parameters β_1 and β_2, the likelihood function for the parameters α_1, α_2, σ^2 and the parameter regression vectors γ_1 and γ_2 in presence of the fixed covariate vector $\mathbf{x}$ and the random factor w_i based on the ith multivariate observation $(t_{1i}, t_{2i}, \delta_{1i}, \delta_{2i})$ is given, by,

$$\begin{aligned} L(\alpha_1, \alpha_2, \gamma_1, \gamma_2, w_i, \sigma^2) &= \left[1 - \exp\left\{-\left(\frac{t_{1i}}{\beta_{1i}}\right)^{\alpha_1}\right\}\right]^{1-\delta_{1i}} \left[\frac{\alpha_1}{\beta_{1i}^{\alpha_1}} t_{1i}^{\alpha_1 - 1} \exp\left\{-\left(\frac{t_{1i}}{\beta_{1i}}\right)^{\alpha_1}\right\}\right]^{\delta_{1i}} \\ &\times \left[1 - \exp\left\{-\left(\frac{t_{2i}}{\beta_{2i}}\right)^{\alpha_2}\right\}\right]^{1-\delta_{2i}} \left[\frac{\alpha_2}{\beta_{2i}^{\alpha_2}} t_{2i}^{\alpha_2 - 1} \exp\left\{-\left(\frac{t_{2i}}{\beta_{2i}}\right)^{\alpha_1}\right\}\right]^{\delta_{2i}} \end{aligned} \tag{3.4}$$

Inferences for the parameters $\alpha_1, \alpha_2, \gamma_1, \gamma_2$ and $\tau = 1/\sigma^2$ are obtained using a Bayesian hierarchical approach in two stages. We assume Gamma(a, b) *prior* distributions for the parameters α_1, α_2 with a and b known hyperparameters, and Gamma(a, b) denotes a Gamma distribution with mean a/b and variance a/b^2; and $N(c, d^2)$ *prior* distributions for the regression parameters $\gamma_{j0}, \gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jp}, j = 1, 2$ in the first stage of the hierarchical Bayesian approach; in the second stage, we assume a Gamma *prior* distribution for the parameter $\tau = 1/\sigma^2$ associated to the Normal distribution $N(0, \sigma^2)$ assumed for the random factors $w_i, i = 1, 2, \ldots, n$. Let us denote this model as "model 1".

Tobit models for left-censored data

Another possibility in the data analysis in presence of left-censored data is to consider a Tobit model [Tobin, 1958], that could fit the data by assuming a regression model whose response variable is censored to a prefixed limiting value. The censoring occurs when the response of the regression model is not directly observable, but its independent variables (or covariates) are observed. Tobit models usually assumes the normality assumption but could be modeled by other probability distributions (see, for example, [Martínez-Flórez et al., 2013]).

If we have a complete observation, that is, $(T > C)$, let us assume a truncated Weibull distribution with probability density function given by,

$$f(t \mid T > C) = \frac{f_0(t)}{P(T > C)} \tag{3.5}$$

where $f_0(t) = \alpha/\beta^\alpha t^{\alpha-1} \exp\{-(t/\beta)^\alpha\}$ and $S_0(t) = P(T > t) = \exp\{-(t/\beta)^\alpha\}$. In this way, let us assume the mixture model, given by the probability density function,

$$f(t) = p\delta C(t) + (1-p)\frac{f_0(t)}{S_0(C)} \tag{3.6}$$

where $\delta C(t)$ is the Dirac measure at C and p is the associated probability of T to be left-censored for the mixture model and 1-p is the probability to be non-censored data. In this case, if $T \leq C$, S(t) = 1; otherwise, if $T > C$, $S(t) = (1-p)S_0(t)/S_0(C)$ where $S_0(C) = \exp\{-(C/\beta)^\alpha\}$. Observe that for this truncated mixture model the expected value for $T > C$, is given by $E(T) = (1-p)\beta\Gamma(1+1/\alpha)/S_0(C)$ where C is fixed (left-censoring). The likelihood function for the parameters p, α and β based on the i-th observation is given by,

$$L(p, \alpha, \beta; t_i) = p\delta C(t_i) + (1-p)\frac{f_0(t_i)}{S_0(C)} \tag{3.7}$$

With the censoring information, let us define a binary variable $\delta = 1$ if T is a complete observation $(T > C)$ and $\delta = 0$ if T is a left censored observation $(T \leq C)$ with conditional probabilities given by

$$\begin{aligned} P(\delta = 0 \mid p, \alpha, \beta, t) &= \frac{p}{p + (1-p)\dfrac{f_0(t)}{S_0(C)}} \\ P(\delta = 1 \mid p, \alpha, \beta, t) &= \frac{(1-p)\dfrac{f_0(t)}{S_0(C)}}{p + (1-p)\dfrac{f_0(t)}{S_0(C)}} \end{aligned} \tag{3.8}$$

In this way, we have a Bernoulli distribution where $\delta = 1$ $(T > C)$ or $\delta = 0$ $(T \leq C)$. Thus, the likelihood function $L(p, \alpha, \beta)$ based on n observations is given by

$$L(p, \alpha, \beta; \mathbf{t}, \boldsymbol{\delta}) = \prod_{i=1}^{n} p^{(1-\delta_i)} \left[(1-p)\frac{f_0(t_i)}{S_0(C)}\right]^{\delta_i} \tag{3.9}$$

For our analysis, we assume a truncated Weibull distribution. Moreover, in the analysis of bivariate data in presence of a covariate vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)^\top$ affecting both dependent random variables T_1 and T_2, we also assume Weibull distributions Wei(α_1, β_1) and Wei(α_2, β_2), respectively, as considered in section (3.2). In this way, we assume the

same regression models for the scale parameters β_j given by (2.3) and logistic models for the parameters p_{ji}, given by,

$$logit(p_{ji}) = \log\left(\frac{p_{ji}}{1-p_{ji}}\right) = \zeta_{j0} + \zeta_{j1}x_{1i} + \zeta_{j2}x_{2i} + \ldots + \zeta_{jp}x_{pi} + w_i \tag{3.10}$$

for $j = 1, 2; i = 1, 2, \ldots, n$. Observe that we are assuming the same random factor w_i assuming a Normal distribution $N(0, \sigma^2)$ to capture the possible dependence between the two responses. Furthermore, the likelihood function for the parameters $\boldsymbol{\alpha_1, \alpha_2, \gamma_1, \gamma_2, \zeta_1}$ and $\boldsymbol{\zeta_2}$, where $\boldsymbol{\gamma_1} = (\gamma_{10}, \gamma_{11}, \gamma_{12}, \ldots, \gamma_{1p})^\top$, $\boldsymbol{\gamma_2} = (\gamma_{20}, \gamma_{21}, \gamma_{22}, \ldots, \gamma_{2p})^\top$, $\boldsymbol{\zeta_1} = (\zeta_{10}, \zeta_{11}, \zeta_{12}, \ldots, \zeta_{1p})^\top$, $\boldsymbol{\zeta_2} = (\zeta_{20}, \zeta_{21}, \zeta_{22}, \ldots, \zeta_{2p})^\top$, assuming different left censoring C_i, based on n observations is given, by,

$$L(\boldsymbol{\alpha_1, \alpha_2, \gamma_1, \gamma_2, \zeta_1, \zeta_2}) = \prod_{i=1}^{n} p_{1i}^{(1-\delta_{1i})}\left[(1-p_{1i})\frac{f_0(t_{1i})}{S_0(C_{1i})}\right]^{\delta_{1i}} \prod_{i=1}^{n} p_{2i}^{(1-\delta_{2i})}\left[(1-p_{2i})\frac{f_0(t_{2i})}{S_0(C_{2i})}\right]^{\delta_{2i}} \tag{3.11}$$

where $\delta_{1i} = 1$ $(T_{1i} > C_{1i})$ or $\delta_{1i} = 0$ $(T_{1i} \leq C_{1i})$ and $\delta_{2i} = 1$ $(T_{2i} > C_{2i})$ or $\delta_{2i} = 0$ $(T_{2i} \leq C_{2i})$. For some applications, we could have same fixed left censoring values in (3.11), that is, C_1 and C_2 in place of C_{1i} and C_{2i}.

For a hierarchical Bayesian analysis of the model, we assume Gamma(a, b) *prior* distributions for the parameters α_1 and α_2 and Normal(c, d^2) *prior* distributions for the regression parameters $\gamma_{10}, \gamma_{11}, \ldots, \gamma_{1p}; \gamma_{20}, \gamma_{21}, \ldots, \gamma_{2p}; \zeta_{10}, \zeta_{11}, \ldots, \zeta_{1p}$ and $\zeta_{20}, \zeta_{21}, \ldots, \zeta_{2p}$ with a, b, c and d known hyperparameters in the first stage of the hierarchical Bayesian analysis. In the second stage of the hierarchical Bayesian analysis we assume the same gamma *prior* for the parameter $\tau = 1/\sigma^2$ assumed in "model 1". Let us denote this model, as "model 2". We use MCMC simulation methods, Metropolis-within-Gibbs algorithms, to get *posterior* summaries of interest for the parameters of the models introduced in Sections (3.2) and (3.2) (see, for example, Chib and Greenberg, 1995, Gelfand and Smith, 1990, Gelman et al., 2013, Gilks et al., 1996).

3.3 Application to a Stellar Astronomy Dataset

Classical approach assuming standard polynomial regression models

First of all, we assume a preliminary data analysis of the astronomy data introduced in Appendix 7, assuming the responses abundance of beryllium (Be) and lithium (Li) as two independent random variables in presence of two covariates Type (Type = 1 indicates planet-hosting stars and Type = 2 is the control sample) and Teff (in degrees Kelvin) is

the stellar surface temperature not considering the presence of the censored data ($n = 55$ uncensored observations for the response Be and $n = 36$ uncensored observations for the response Li). Figure 2 shows the scatterplots of the response abundance of beryllium (Be) and the response lithium (Li) versus Type and Teff in the logarithm scale.

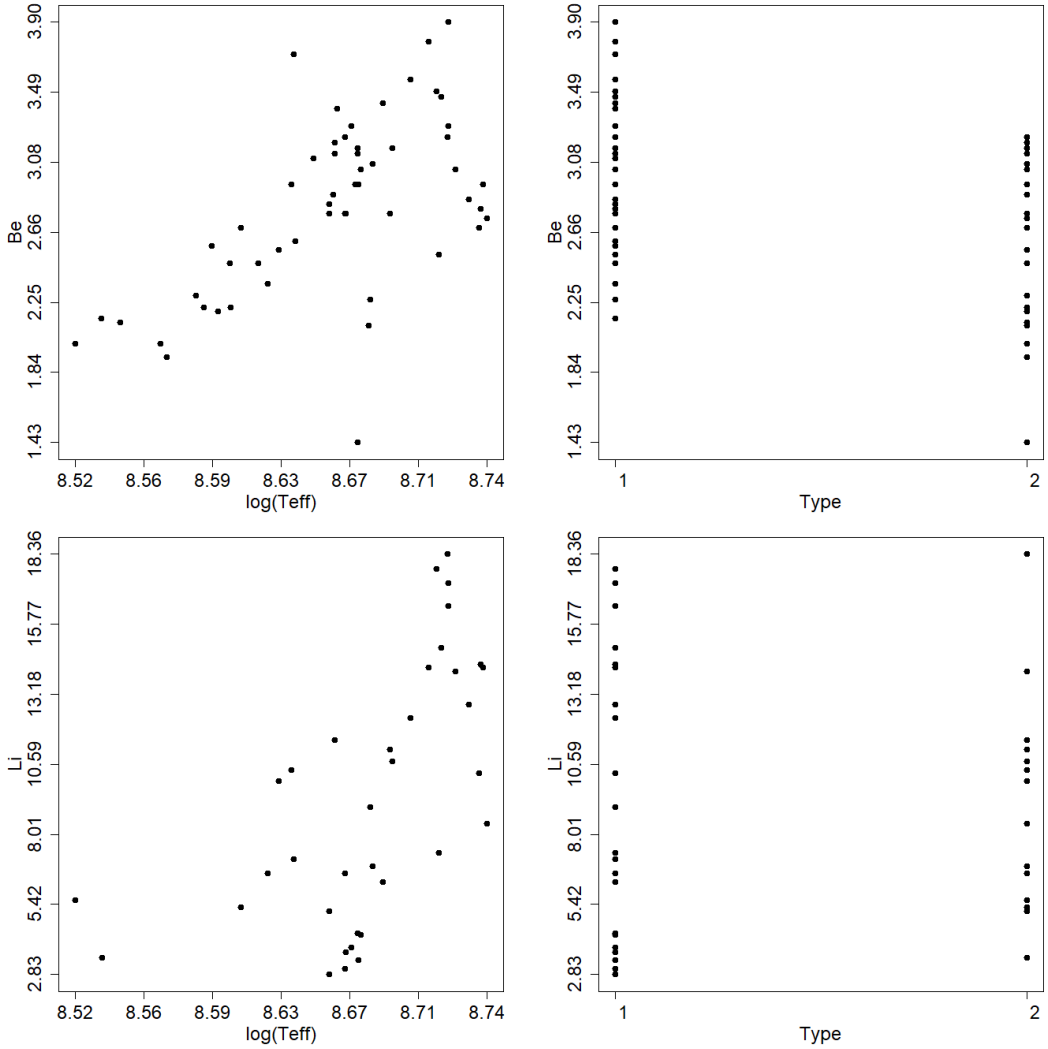

Figure 2 – Scatterplots of Be (upper panels) and Li (lower panels) versus Type and log(Teff).

From the plots of Figure 2, we observe that the response abundance of beryllium (Be) is smaller with the control sample (Type = 2) when compared to planet-hosting stars and increases with larger stellar surface temperature Teff in the logarithm scale. We also observe that the response lithium (Li) is similar with the control sample (Type = 2) with the planet-hosting stars and increases with larger stellar surface temperature Teff in the logarithm scale. From Figure 2, also it is observed the presence of possible curvature for the relation of both tesponses Be and Li versus log(Teff). Assuming polynomial regression

models of order three with standard normal errors (linear, quadratic and cubic effects) for the responses abundance of beryllium (Be) and lithium (Li) in logarithm scale in presence of the covariate log(Teff), Table 7 shows the least square estimators (LSE) for the regression parameters of the polynomial regression models. The needed assumptions for the polynomial regression models were reasonably verified from residual plots.

Table 7 – LSE for the polynomial regression models with the covariate log(Teff).

Source	DF	SS	F	p-value
Response: log(Be)				
Linear	1	0.715987	28.29	< 0.001
Quadratic	1	0.061542	2.50	0.120
Cubic	1	0.048961	2.03	0.160
Response: log(Li)				
Linear	1	3.75767	17.34	< 0.001
Quadratic	1	1.06753	5.59	0.024
Cubic	1	0.02557	0.13	0.720

From the obtained results of Table 7, we see that assuming a significance level equal to 5%, the linear effect of log(Teff) is significative (p-value < 0.05) in the response log(Be); the linear and quadratic effects of log(Teff) are significative (p-value < 0.05) in the response log(Li).

A hierarchical Bayesian analysis assuming the bivariate data in the original scale and left-censoring

In this section, we assume dependent responses abundance of beryllium (Be) and lithium (Li) in presence of the two covariates Type (Type = 1 indicates planet-hosting stars and Type = 2 is the control sample) and Teff (in degrees Kelvin), the stellar surface temperature, considering all dataset presented in Appendix 7, that is, $n = 66$ observations, including the non-censored and the left-censored data in the original scale. We assume Weibull distributions Wei(α_1, β_1) and Wei(α_2, β_2), for the two responses Be and Li with regression models 3.3 for the scale parameters in presence of the covariates Type and Teff and a random factor W which captures the possible dependence between Be and Li under a hierarchical Bayesian analysis. That is, we assume the regression models given by,

$$\begin{aligned} \beta_{1i} &= \exp(\gamma_{10} + \gamma_{11} type_i + \gamma_{12} \log(T_{eff})_i + w_i) \\ \beta_{2i} &= \exp(\gamma_{20} + \gamma_{21} type_i + \gamma_{22} \log(T_{eff})_i + \gamma_{23} [\log(T_{eff})_i]^2 + w_i) \end{aligned} \tag{3.12}$$

where, $i = 1, 2, ..., 66$; w_i is a random factor which captures extra-Weibull variability and possible dependence between both dependent variables assumed to be independent random variables with a Normal$(0, \sigma^2)$ distribution. The inclusion of the factors type, log(Teff) and [log(Teff)]^2 in the regression models for β_1 and β_2 (3.3) was based from the obtained results in Section (3.2).

For a Bayesian analysis, we assume uniform *prior* distributions $U(0, 10)$ for the parameters α_1 and α_2; U(0, 200) for the parameter $\tau = 1/\sigma^2$; N(0,1) for the parameters $\gamma_{11}, \gamma_{12}, \gamma_{21}$ and γ_{22}; N(0,10) for the parameter γ_{23}; and N(0, 100) for the parameters γ_{10} and γ_{20}. That is, we are assuming approximately non-informative *prior* distributions for all parameters. We further assume *prior* independence among the parameters. Inferences for the parameters of the regression models (3.12) are obtained under a hierarchical Bayesian approach using existing MCMC methods like the Gibbs and the Metropolis-Hastings algorithms.

In the simulation of samples of the joint *posterior* distribution, $\pi(\boldsymbol{\theta}/data)$ where $\boldsymbol{\theta}$ is the vector of all parameters, we use Gibbs or Metropolis-Hastings algorithms (Gelfand and Smith, 1990, Chib and Greenberg, 1995), where it is needed to sample each parameter from the *posterior* conditional distributions $\pi(\theta_r/\boldsymbol{\theta}(r), data)$, where $\boldsymbol{\theta}(r)$ denotes the vector of all parameters except θ_r and r is associated to each one of the parameters of the model. In this study, we use the OpenBugs software [Spiegelhalter et al., 2003] in the simulation of samples of the joint *posterior* distribution of interest which simplifies the computational work, since this software only requires the definition of the likelihood function for $\boldsymbol{\theta}$ and the *prior* distribution $\pi(\boldsymbol{\theta})$.

A burn-in sample of size 111,000 was deleted to eliminate the effects of the initial values in the iterative simulation process and a final Gibbs sample of size 1000 (taken every 100th simulated Gibbs sample) was used to get the *posterior* summaries of interest. Convergence of the simulation algorithm was verified from trace plots of the simulated Gibbs samples. Table 4 shows the *posterior* means, *posterior* standard-deviations and 95% credible intervals for all parameters of the regression models (OpenBugs code in Appendix 7).

Table 8 shows that the stellar surface temperature Teff (in degrees Kelvin) in logarithmic scale, that is, log(Teff), has a significative effect on the response abundance of beryllium (Be) since zero is not included in the 95% credible interval for γ_{12}; the square of the stellar surface temperature Teff (in degrees Kelvin) in logarithmic scale (quadratic effect), that is, [log(Teff)$_i$]^2 , has a significative effect on the response abundance of lithium

(Li) since zero is not included in the 95% credible interval for γ_{23}. All other covariates do not show significative effects on the responses Be and Li since zero is included in the credible intervals for the correspoding regression parameters. Figure 3 shows the residual plots of the fitted proposed bivariate Weibull regression model.

Table 8 – *Posterior* summaries for the bivariate Weibull regression model (model 1).

Parameter	Mean	Std. Dev.	95% Cred. Int.	
			Lower	Upper
α_1	5.0491	0.7201	3.7900	6.5141
α_2	0.9637	0.1435	0.6997	1.2640
γ_{10}	-14.8623	3.9210	-22.8901	-6.7581
γ_{11}	-0.0851	0.0573	-0.2046	0.0295
γ_{12}	1.8491	0.4514	0.9307	2.7830
γ_{20}	-25.1201	7.5720	-39.7910	-10.2712
γ_{21}	0.0734	0.2851	-0.5058	0.6497
γ_{22}	-0.8785	0.9969	-2.8601	1.0720
γ_{23}	0.4554	0.1263	0.1981	0.6973
τ	149.9001	36.8422	67.3620	197.7101

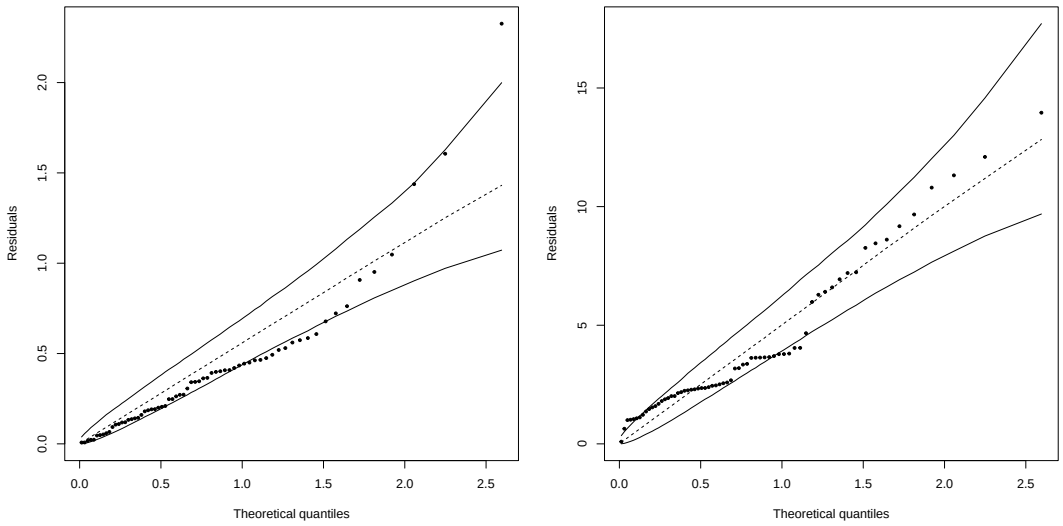

Figure 3 – Residuals of the fitted proposed bivariate Weibull regression model for the responses Be (left panel) and Li (right panel).

A hierarchical Bayesian analysis for the Tobit-Weibull model assuming the bivariate data

As an alternative model, in this section we also assume the dependent responses abundance of beryllium (Be) and lithium (Li) in the original scale with Weibull distributions Wei(α_1, β_1) and Wei(α_2, β_2), respectively, in presence of the two covariates Type (Type

= 1 indicates planet-hosting stars and Type = 2 is the control sample) and Teff (in degrees Kelvin), the stellar surface temperature, considering now the Tobit-Weibull model introduced in Section 3 given by the following regression models,

$$\begin{aligned} \beta_{1i} &= \exp(\gamma_{10} + \gamma_{11} type_i + \gamma_{12} \log(T_{eff})_i) + w_i \\ \beta_{2i} &= \exp(\gamma_{20} + \gamma_{21} type_i + \gamma_{22} \log(T_{eff})_i + \gamma_{23} [\log(T_{eff})_i]^2) + w_i \end{aligned} \tag{3.13}$$

and,

$$\begin{aligned} logit(p_{1i}) = \log\left(\frac{p_{1i}}{1-p_{1i}}\right) = \zeta_{10} + \zeta_{11} type_i + \zeta_{12} \log(T_{eff})_i + w_i \\ logit(p_{2i}) = \log\left(\frac{p_{2i}}{1-p_{2i}}\right) = \zeta_{20} + \zeta_{21} type_i + \zeta_{22} \log(T_{eff})_i \zeta_{23} [\log(T_{eff})_i]^2) + w_i \end{aligned} \tag{3.14}$$

where $i = 1, 2, \ldots, n$ (sample size); w_i is a random factor which captures extra-Weibull variability and dependence between both dependent variables assumed to be independent random variables with a N(0, σ^2) distribution.

For a Bayesian analysis, we assume Gamma *prior* distributions G(1,1) for the parameters α_1 and α_2; U(0,100) for the parameter $\tau = 1/\sigma^2$; $N(0, 0.01)$ for the parameters $\gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}$ and γ_{23}; $N(0, 1)$ for the parameters γ_{10} and γ_{20}; N(0,0.01) for the parameters $\zeta_{11}, \zeta_{12}, \zeta_{21}, \zeta_{22}$ and ζ_{23}; N(0,1) for the parameters ζ_{10} and ζ_{20}. We further assume *prior* independence among the parameters. Inferences for the parameters of the regression models above are also obtained under a hierarchical Bayesian approach using existing MCMC methods, as Metropolis-within-Gibbs algorithms.

A burn-in sample of size 11,000 was deleted to eliminate the effects of the initial values in the iterative simulation process and a final Gibbs sample of size 2000 (taking every 100th simulated Gibbs sample) was used to get the *posterior* summaries of interest. Table 3 shows the *posterior* means, *posterior* standard-deviations and 95% credible intervals for all parameters of the regression models (OpenBugs code in Appendix 7). Figure 4 shows the residual plots of the fitted Tobit-Weibull proposed model.

Table 9 also shows that using models (3.13) and (3.14), the covariates Type (Type = 1 indicates planet-hosting stars and Type = 2 is the control sample) and the stellar surface temperature Teff (in degrees Kelvin) in logarithmic scale, that is, log(Teff), have a significative effect on the scale parameter of the Weibull distribution assumed for the response abundance of beryllium (Be) since zero is not included in the 95% credible intervals for γ_{11} and γ_{12}; the square of the stellar surface temperature Teff (in degrees Kelvin) in logarithmic scale (quadratic effect), that is, $[\log(\text{Teff})_i]^2$, has a significative effect on the scale parameter of the Weibull distribution assumed for the response abundance

of lithium (Li) since zero is not included in the 95% credible interval for γ_{23}. All other covariates do not show significative effects associated to the responses Be and Li since zero is included in the credible intervals for the corresponding regression parameters. Figure 4 shows the residual plots of the fitted bivariate Tobit-Weibull model.

Table 9 – *Posterior* summaries for the Tobit-Weibull model (model 2).

Parameter	Mean	Std. Dev.	95% Cred. Int.	
			Lower	Upper
α_1	10.3869	1.8797	7.4579	15.1100
α_2	2.6570	0.2655	2.1640	3.2310
γ_{10}	-0.5008	0.3841	-1.2970	0.0984
γ_{11}	-0.1282	0.0391	-0.2030	-0.0446
γ_{12}	0.2042	0.0431	0.1351	0.2896
γ_{20}	-0.6091	1.0391	-2.2260	1.5081
γ_{21}	-0.0348	0.0725	-0.1817	0.1054
γ_{22}	-0.0592	0.1038	-0.2782	0.1198
γ_{23}	0.0467	0.0203	0.0151	0.0833
τ	66.6480	18.0801	33.4101	98.4501
ζ_{10}	-1.1667	0.6690	-2.4241	0.1544
ζ_{11}	-0.0271	0.0959	-0.2268	0.1633
ζ_{12}	-0.1162	0.0771	-0.2711	0.0333
ζ_{20}	0.4024	0.9918	-1.6111	2.2770
ζ_{21}	0.0429	0.0969	-0.1466	0.2343
ζ_{22}	0.0202	0.1009	-0.1756	0.2076
ζ_{23}	-0.0347	0.0175	-0.0687	0.0012

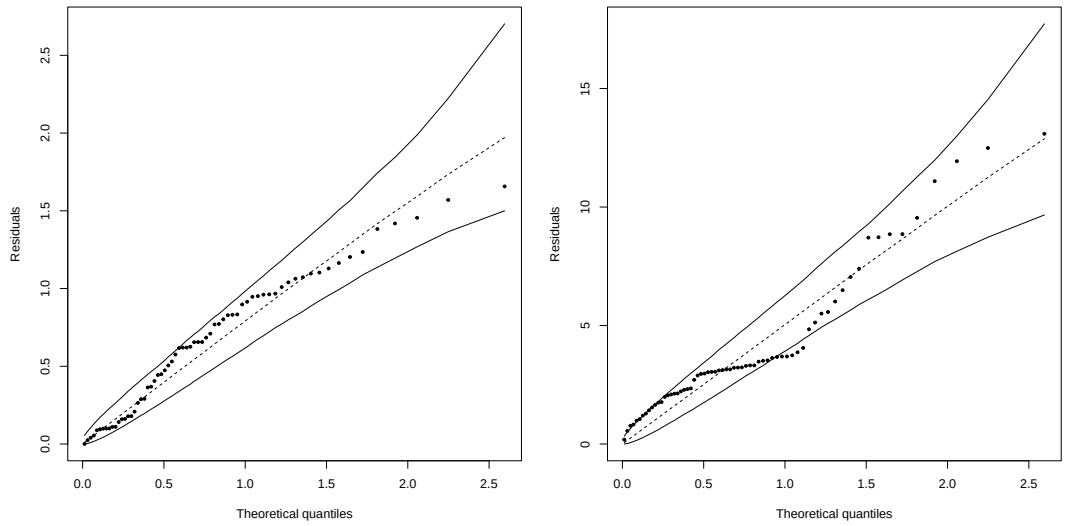

Figure 4 – Residuals of the fitted proposed bivariate Tobit-Weibull model for the responses Be (left panel) and Li (right panel).

From the obtained inference results we observe that the covariate $[\log(\text{Teff})_i]$ has significative effect on the responses Be and Li assuming the standard Weibull likelihood model (model 1) (linear effect on Be and quadratic effect on Li). Assuming the Tobit-Weibull likelihood model (model 2) it is observed that the covariate $[\log(\text{Teff})_i]$ has significative effect on the responses Be (linear effect) and Li (quadratic effect) and the covariate Type has significative effect on the response Be. Since the residual plots in figures 3 and 4, have very similar fits, we will make a comparison using the estimated expected values assuming models 1 and 2 in the next section.

Some remarks on the fit of models 1 and 2 for the astronomy data

To compare both models (model 1 and model 2), we consider plots of the Monte Carlo Bayesian estimators based on the simulated Gibbs samples of the expected values for both responses Be and Li versus the observed values assuming the fitted models. Figure 5 shows the plots of the expected values $E(T_{ji}) = \beta_i\Gamma(1 + 1/\alpha_j)$ for model 1 and approximated $E(T_{ji}) = (1 - p_{ji})\beta_i\Gamma(1 + 1/\alpha_j)/S_0(C_i)$ for the Tobit-Weibull truncated model 2, $i = 1, 2, \ldots, n; j = 1, 2$ for the responses Be and Li and observed values considering the proposed models 1 and 2, where,

i.) Model 1 for the response Be; the expected mean value is given by:

$$E(T_{1i}) = \hat{\beta}_{1i}\Gamma\left(1 + \frac{1}{5.0491}\right)$$

where $\hat{\beta}_{1i} = \exp\{-14.8623 - 0.0851type_i + 1.8491\log(T_{eff})_i\}$.

ii.) Model 1 for the response Li; the expected mean value is given by:

$$E(T_{2i}) = \hat{\beta}_{2i}\Gamma\left(1 + \frac{1}{0.9637}\right)$$

where $\hat{\beta}_{2i} = \exp\{-25.1201 + 0.0734type_i - 0.8785\log(T_{eff})_i + 0.4554[\log(T_{eff})_i]^2\}$.

iii.) Model 2 for the response Be; we assume as a simplification the plots of the non-censored data, that is, with expected mean value given by $E(T_{ji}) = (1 - p_{ji})\beta_i\Gamma(1 + 1/\alpha_j)$, that is,

$$E(T_{1i}) = (1 - \hat{p}_{1i})\hat{\beta}_{1i}\Gamma\left(1 + \frac{1}{10.38}\right)$$

where $\hat{\beta}_{1i} = \exp\{-0.5008 - 0.1282type_i + 0.2042\log(T_{eff})_i\}$ and $\hat{p}_{1i} = A_{1i}/(1 + A_{1i})$, where $A_{1i} = \exp\{-1.166 - 0.02712type_i - 0.1162\log(T_{eff})_i\}$.

iv.) Model 2 for the response Li; the expected mean value is given by:

$$E(T_{2i}) = (1 - \hat{p}_{2i})\hat{\beta}_{2i}\Gamma\left(1 + \frac{1}{2.65}\right)$$

where $\hat{\beta}_{2i} = \exp\{-0.6091 - 0.03487type_i - 0.0592\log(T_{eff})_i + 0.04675[\log(T_{eff})_i]^2\}$ and $\hat{p}_{2i} = A_{2i}/(1+A_{2i})$, where $A_{2i} = \exp\{0.4024 + 0.04295type_i + 0.02027\log(T_{eff})_i - 0.03474[\log(T_{eff})_i]^2\}$.

From the plots of Figure 5, we observe that model 1 gives, in general, estimated expected means closer to the observed data Be an Li, indicating better fit of model 1 for the astronomy data when compared to model 2.

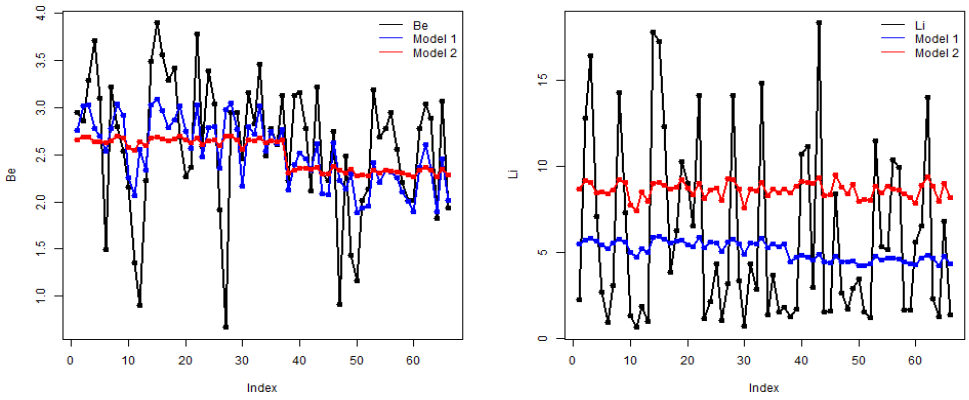

Figure 5 – Plots of the expected values and the responses Be and Li assuming model 1 and model 2.

3.4 Concluding remarks

In the application with the astronomy data, we observed that the obtained Bayesian inference results lead to similar results considering both proposed models, in terms of discovering the significative effects of the covariates Type (Type = 1 indicates planet-hosting stars and Type = 2 is the control sample) and Teff (in degrees Kelvin) is the stellar surface temperature on both astronomy responses abundance of beryllium (Be) and lithium (Li) and with similar computational costs to simulate samples for the joint *posterior* distributions of interest using the free OpenBugs software.

In the application considered in this study, we observed that assuming the standard Weibull likelihood approach, we obtained better model fit for the data (see Figure 4). Other applications and possibly, some simulation studies, should be considered in future

studies to compare the adequability and performance of the two proposed models (models 1 and 2) in each application.

The great advantage of the proposed hierarchical Bayesian methodology in the analysis of bivariate data is the simple form of the likelihood given by product of the likelihood functions and the dependence structure given by a non-observed latent factor or frailty which also could be generalized to other structures.

Other point, especially in applications, in favor of our approach: the use of parametric bivariate probability models derived from copula functions, usually depends on the choice of a particular copula function among hundreds of existing copula functions, since each copula represents different dependence structure for the dataset. It is also interesting to point out that the Tobit model gives better interpretations of interest to researchers. Usually, mixture models as considered in the Tobit model given by 3.7, have some advantages in the interpretations, in the same way as obtained with the use of cure fraction models where it is possible to get estimator for susceptible and non-susceptible individuals that can die from some diseases (Maller and Zhou, 1996, Achcar et al., 2012, de Oliveira et al., 2019).

In addition, other existing parametric lifetime distributions as exponential, gamma, log-normal or generalizations of the Weibull distribution could be considered to model the univariate distributions for the two responses of the bivariate data in presence of left-censored data. Finally, it is important to point out that the use of existing Bayesian simulation softwares like the OpenBugs software leads to great simplification in obtaining the Bayesian inferences of interest. Another advantage of the Bayesian methodology: it is possible to use expert opinion in the elicitation of *prior* distributions that can lead to more accurate inference results.

Chapter 4

A Bayesian approch for univariate or bivariate lifetime data in presence of left-censored data assuming a Weibull-Tobit model

4.1 Introduction

In many applications, especially in medical or engineering studies we could have two lifetimes associated to the same individual. In some cases, these two lifetimes are assumed to be independent, but the lifetime of one component could influence the lifetime of the other component in which case it becomes necessary to introduce a dependence structure between the two variables. This is the case, for example, considering the failure times of paired organs like kidney, lungs, eyes, ears, dental implants among many others. To analyse bivariate lifetimes we could assume different parametric distributions introduced in the literature (see Freund, 1961, Marshall and Olkin, 1985, Gumbel, 1960, Hawkes, 1972, Hougaard, 1986, Arnold and Strauss, 1988) or to use Bayesian hierarchical methods. The main goal of this paper is to introduce a hierarchical Bayesian analysis for bivariate lifetimes assuming Tobit-Weibull models for their marginal distributions in presence of left-censoring mechanism. The possible dependence structure between the bivariate data is modeled by the introduction of a frailty or latent variable. The main reason for the use of the Weibull distribution, usually the most used lifetime distribution in lifetime data applications is due to the great flexibility of fit for the data. Besides the great flexibility of fit, the Weibull distribution usually assumed in lifetime data analysis has only two parameters, which implies in great simplicity to get the inferences of interest, especially assuming a left-censored scheme.

In this study, we to introduce the univariate and bivariate models based on the Tobit-Weibull distribution as an alternative to left-censored data analysis. The Chapter is organized as follows: Section 4.2 presents the proposed Tobit-Weibull model approach for univariate and bivariate data assuming data with left-censoring mechanism and some inference methods for the parameters of the model. Section 4.3 presents medical applications of the proposed methodology under a hierarchical Bayesian approach. Finally, Section 4.4 closes the paper with some concluding remarks and directions for future research.

4.2 Materials and Methods

Univariate Tobit-Weibull Model

Suppose Y is a random variable denoting the lifetime of an unit or patient such that the lifetime data is given by $T = \max(C, Y)$ where C is a censored time and Y is a complete observation. Thus, we could define a indicator variable as,

$$\delta = \begin{cases} 1, \text{ if T is a complete observation } (Y > C) \\ 0, \text{ if T is a left censored observation } (Y \leq C) \end{cases} \tag{4.1}$$

Notice that $\delta = 1$ if T is a complete observation $(Y > C)$ and $\delta = 0$ if T is a left censored observation $(Y \leq C)$. If we have a complete observation, that is, $(Y > C)$, let us assume a truncated Weibull distribution with probability density function given by,

$$f(t \mid T > C) = \frac{\frac{\alpha}{\beta^\alpha} t^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\}}{\exp\left\{-\left(\frac{C}{\beta}\right)^\alpha\right\}} \tag{4.2}$$

Using the mixing approach, the density function of Tobit-Weibull model is given by the equation,

$$f(t) = p\delta_c(t) + (1-p)\frac{\frac{\alpha}{\beta^\alpha} t^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\}}{\exp\left\{-\left(\frac{C}{\beta}\right)^\alpha\right\}} \tag{4.3}$$

where $\delta_c(t)$ is the Dirac measure at $(0, C)$, that is, $\delta_c(t) = 1/C$ if $0 < t < C$ and $\delta_c(t) = 0$ if $t > C$, which guarantees that $f(t)$ is a probability density function, p is the associated probability of T to be left-censored for the mixture model and $1 - p$ is the probability to be non-censored. In this way, we have:

- If $T \leq C$, $S(t) = 1$
- If $T > C$, $S(t) = (1-p)\frac{S_0(t)}{S_0(C)}$

where $S_0(t) = \exp\left\{-\left(\frac{t}{\beta}\right)^{\alpha}\right\}$ and $S_0(C) = \exp\left\{-\left(\frac{C}{\beta}\right)^{\alpha}\right\}$.

In terms of statistical inference, the likelihood function for the parameters p, α and β based on one observation is given from (4.3) by,

$$L(p, \alpha, \beta) = p\delta_c(t) + (1-p)\frac{f_0(t)}{S_0(C)} \tag{4.4}$$

where $f_0(t)$ and $S_0(t)$ are defined by,

$$f_0(t) = \frac{\alpha}{\beta^{\alpha}} t^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^{\alpha}\right\} \quad \text{and} \quad S_0(t) = P(T > t) = \exp\left\{-\left(\frac{t}{\beta}\right)^{\alpha}\right\}$$

With the censoring information (4.1), we get the conditional probabilities,

$$\begin{aligned} P(\delta = 0 \mid p, \alpha, \beta, t) &= \frac{p}{p + (1-p)\frac{f_0(t)}{S_0(C)}} \\ P(\delta = 1 \mid p, \alpha, \beta, t) &= \frac{(1-p)\frac{f_0(t)}{S_0(C)}}{p + (1-p)\frac{f_0(t)}{S_0(C)}} \end{aligned} \tag{4.5}$$

which could be assumed as a Bernoulli trial. In this way, by equations (4.4) and (4.5), the likelihood function based on n observations is reduced to,

$$L(p, \alpha, \beta) = \prod_{i=1}^{n} p^{(1-\delta_i)} \left[(1-p)\frac{f_0(t)}{S_0(C)}\right]^{\delta_i} \tag{4.6}$$

Assuming the truncated Weibull distribution, the likelihood function and the log-likelihood function for the parameters of the Tobit-Weibull model are given respectively by,

$$L(p, \alpha, \beta) = \prod_{i=1}^{n} p^{(1-\delta_i)} \left[(1-p)\frac{\frac{\alpha}{\beta^{\alpha}} t^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^{\alpha}\right\}}{\exp\left\{-\left(\frac{C}{\beta}\right)^{\alpha}\right\}}\right]^{\delta_i} \tag{4.7}$$

and

$$\begin{aligned}\ell(p,\alpha,\beta) &= \sum_{i=1}^{n}\Bigg\{(1-\delta_i)\log(p)+\delta_i\Bigg\{\log(1-p)+\log(\alpha)\\ &+ (\alpha-1)\log(t_i)-\alpha\log(\beta)-\left(\frac{t_i}{\beta}\right)^{\alpha}+\left(\frac{C}{\beta}\right)^{\alpha}\Bigg\}\Bigg\} \qquad (4.8)\end{aligned}$$

Maximum likelihood and Bayesian inference approaches are considered to get the inferences of interest for the proposed model. Maximum likelihood estimators (MLE) for the parameters of the proposed model are obtained using existing numerical optimization procedures as for example, the Gauss-Newton iterative method. The MLE are obtained solving the equations obtained from the first derivatives of the log-likelihood function with respect to the parameters of the model being equal to zero. For a Bayesian approach of the proposed model, we use existing MCMC simulation methods, as Metropolis-within-Gibbs, to get the *posterior* summaries of interest (see, for example, Chib and Greenberg, 1995, Gelfand and Smith, 1990, Gelman et al., 1995, Geman and Geman, 1984, Gilks et al., 1995) assuming independent Gamma(a,b) *prior* distributions for the parameters α and β with a and b known hyperparameters and a Beta(e,f) distribution for p with e and f known hyperparameters.

In presence of covariates, based on this model, we could also introduce the following linear regression structure for a vector of covariates $\mathbf{x}=(x_1,x_2,\ldots,x_p)^\top$ in the scale parameter β given by,

$$\beta=\exp(\gamma_0+\gamma_1x_1+\gamma_2x_2+\ldots+\gamma_px_p) \qquad (4.9)$$

where $\boldsymbol{\gamma}=(\gamma_0,\gamma_1,\gamma_2,\ldots,\gamma_p)^\top$ is the regression parameter vector associated to covariate vector $\mathbf{x}=(x_0,x_1,x_2,\ldots,x_p)^\top$. However, since the mixing parameter is also our target, we assume a logistic regression model for the parameter p given by,

$$\text{logit}(p)=\log\left(\frac{p}{1-p}\right)=\varphi_0+\varphi_1x_1+\varphi_2x_2+\ldots+\varphi_px_p \qquad (4.10)$$

Bivariate Tobit-Weibull Model

For the analysis of bivariate data (T_1,T_2) in presence of a covariate vector $\mathbf{x}=(x_1,x_2,...,x_p)^\top$ affecting both dependent random variables, assume two Weibull distributions, denoted respectively by Wei(α_1,β_1) and Wei(α_2,β_2), for the use of hierarchical

Bayesian methods. Thus, a linear regression structure as considered in the univariate case is assumed for the scale parameters β_j in the Weibull density, that is,

$$\beta_{ji} = \exp(\gamma_{j0} + \gamma_{j1}x_{1i} + \gamma_{j2}x_{2i} + \ldots + \gamma_{jp}x_{pi} + w_i) \tag{4.11}$$

where $\boldsymbol{\gamma_j} = (\gamma_{j0}, \gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jp})^\top$ is the regression parameter vector associated to the covariate vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)^\top$, $j = 1, 2; i = 1, 2, \ldots, n$ (sample size); w_i is a random factor which captures extra-Weibull variability and dependence structure between both dependent variables (T_1, T_2). The random factors or latent variables (non-observed) $W_i, i = 1, \ldots, n$, are assumed to be independent random variables with a Normal $N(0, \sigma^2)$ distribution.

However, our goal is to work with a bivariate Tobit-Weibull model. In this case, we also assume a regression structure based on logistic models for the mixing parameters p_{ji} given by

$$\text{logit}(p_{ji}) = \log\left(\frac{p_{ji}}{1-p_{ji}}\right) = \varphi_{j0} + \varphi_{j1}x_{1i} + \varphi_{j2}x_{2i} + \ldots + \varphi_{jp}x_{pi} + w_i \tag{4.12}$$

for $j = 1, 2; i = 1, 2, \ldots, n$. Observe that we are assuming the same random factor w_i considered for the regression models of the scale parameters with a Normal distribution $N(0, \sigma^2)$ to capture the possible dependence between the two responses. Furthermore, the likelihood function for the parameters $\alpha_1, \alpha_2, \boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\varphi_1}$ and $\boldsymbol{\varphi_2}$, where $\boldsymbol{\gamma_1} = (\gamma_{10}, \gamma_{11}, \gamma_{12}, \ldots, \gamma_{1p})^\top$, $\boldsymbol{\gamma_2} = (\gamma_{20}, \gamma_{21}, \gamma_{22}, \ldots, \gamma_{2p})^\top$, $\boldsymbol{\varphi_1} = (\varphi_{10}, \varphi_{11}, \varphi_{12}, \ldots, \varphi_{1p})^\top$, $\boldsymbol{\varphi_2} = (\varphi_{20}, \varphi_{21}, \varphi_{22}, \ldots, \varphi_{2p})^\top$, assuming different left censoring C_i, based on n observations is given by

$$L(\alpha_1, \alpha_2, \boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\varphi_1}, \boldsymbol{\varphi_2}) = \prod_{i=1}^{n} p_{1i}^{(1-\delta_{1i})}\left[(1-p_{1i})\frac{f_0(t_{1i})}{S_0(C_{1i})}\right]^{\delta_{1i}} \prod_{i=1}^{n} p_{2i}^{(1-\delta_{2i})}\left[(1-p_{2i})\frac{f_0(t_{2i})}{S_0(C_{2i})}\right]^{\delta_{2i}}$$

where $\delta_{1i} = 1(T_{1i} > C_{1i})$ or $\delta_{1i} = 0(T_{1i} \leq C_{1i})$ and $\delta_{2i} = 1(T_{2i} > C_{2i})$ or $\delta_{2i} = 0(T_{2i} \leq C_{2i})$. For some applications, we could have same fixed left censoring values in (12), that is, C_1 and C_2 in place of C_{1i} and C_{2i}.

4.3 Application to health datasets

4.3.1 Thyroid Cancer Data

In this application we consider a left-censored medical data related to differentiated thyroid cancer, fitting the Tobit-Weibull models in two ways: assuming the data as univariate independent data or assuming bivariate dependent data in presence of three

covariates: size, sex and persistence. For instance, the dataset consists in 91 patients and it was also used López et al. [2014] in a descriptive study to evaluate the relationship between the initial thyroglobulin levels and the presence of recurrence of cancer one year after receiving treatment (this dataset is presented in Appendix 7). Basically, each patient received surgery to remove the thyroid gland and then they were treated with radioiodine I-131. The random variables of interest for our goal are: immediately before starting therapy with iodine, a sample of blood was obtained to measure the thyroglobulin level (T_1) of each patient; and after receiving the therapy, the measure of the thyroglobulin level was reported approximately a year after the last therapy session (T_2). The information about thyroglobulin levels data are left censored, since the measuring instrument does not detect values less than 0.1. Also the study considered three covariates which coud be related to the two responses of interest: sex (male=0; female =1); size measures as millimeter (if size $< 40mm = 0$ and size $\geq 40mm = 1$) and persistence measures as nanograms per milliliter (if $TG < 2ng/ml = 0$ and $TG \geq 2ng/ml = 1$).

The statistical analysis was carried out in the `R` software (R Core Team, 2015) and the `R2jags` package was used to obtain the Bayesian estimates for the model parameters. The computer code is presented in Appendix 3.2. In a Bayesian framework, one may assume that no specialized information is available to justify the choice of non-informative *prior* distributions for the model parameters. In this context, we specify *prior* distributions such that, even for moderate sample sizes, the information provided by the data should dominate the *prior* information. The non-informative *prior* distributions adopted in this work are given by, $\alpha_j \sim \text{Gamma}(0.001, 0.001)$, $\boldsymbol{\varphi}_j \sim \text{Normal}_{q+1}(\mathbf{0}, 10^2\boldsymbol{\mathcal{I}}_q)$ and $\gamma_j \sim \text{Normal}_{q+1}(\mathbf{0}, 10^2\boldsymbol{\mathcal{I}}_q)$ where $\boldsymbol{\mathcal{I}}_{q+1}$ is a identity matrix of size $q+1$. The results for each fitted model are presented in Table 10.

Based on the obtained results assuming the Bayesian estimates for the parameters of the independent univariate Tobit-Weibull models presented in Table 10, we observe from the obtained 95% credible intervals (the zero value are not inside the intervals), that the thyroglobulin level before starting therapy ($T1$) is affected by size of tumor with positive regression parameter estimative 0.0645 considering the linear structure and negative regression parameter estimative -0.27009 considering the logistic structure. We also observe that the thyroglobulin level after the last therapy session ($T2$) is affected by the covariates sex (negative parameter regression estimative -1.5462) and persistence (positive parameter regression estimative 4.1276) assuming the linear regression model and persistence (negative parameter regression estimative -1.9554) assuming the logistic regression model.

Table 10 – Summary of the fitted Tobit-Weibull models for thyroid cancer data.

Approach	Parameter	Estimate	Std. Dev.	95% Cred. Int.	
				Lower	Upper
Univariate (T_1)	α	0.2785	0.0296	0.2216	0.3371
	γ_0 (Intercept)	0.0011	0.0995	-0.1941	0.1960
	γ_1 (Sex)	-0.0352	0.0989	-0.2293	0.1586
	γ_2 (Size)	0.0645	0.0195	0.0243	0.1010
	γ_3 (Persistence)	0.0623	0.0998	-0.1326	0.2588
	φ_0 (Intercept)	-0.0389	0.0989	-0.2322	0.1554
	φ_1 (Sex)	-0.0364	0.0992	-0.2310	0.1577
	φ_2 (Size)	-0.2709	0.0547	-0.3884	-0.1751
	φ_3 (Persistence)	-0.0015	0.1001	-0.1987	0.1945
Univariate (T_2)	α	0.6062	0.0867	0.4448	0.7849
	γ_0 (Intercept)	-0.9599	0.5925	-2.1692	0.1567
	γ_1 (Sex)	-1.5462	0.4722	-2.4907	-0.6313
	γ_2 (Size)	0.0164	0.0138	-0.0091	0.0456
	γ_3 (Persistence)	4.1276	0.4458	3.2435	5.0016
	φ_0 (Intercept)	0.2316	0.5526	-0.8540	1.3133
	φ_1 (Sex)	0.0385	0.4979	-0.9311	1.0241
	φ_2 (Size)	-0.0260	0.0180	-0.0629	0.0076
	φ_3 (Persistence)	-1.9554	0.6518	-3.3015	-0.7411
Bivariate (T_1, T_2)	α_1	0.99061	0.02158	0.94095	1.02923
	α_2	0.89352	0.09215	0.70891	1.06788
	γ_{10} (Intercept)	-1.34905	0.42210	-2.26651	-0.56162
	γ_{11} (Sex)	-0.51200	0.43559	-1.31960	0.45562
	γ_{12} (Size)	0.03684	0.53783	-0.99534	1.10657
	γ_{13} (Persistence)	-0.27181	0.69470	-1.67177	1.11800
	γ_{20} (Intercept)	1.54971	0.54668	0.44780	2.58645
	γ_{21} (Sex)	0.85854	0.55499	-0.29043	1.90740
	γ_{22} (Size)	-0.03191	0.67645	-1.39659	1.23435
	γ_{23} (Persistence)	-0.10521	0.73191	-1.51705	1.34707
	φ_{10} (Intercept)	-1.71107	0.74075	-3.17986	-0.28288
	φ_{11} (Sex)	-1.13751	0.80809	-2.75314	0.44813
	φ_{12} (Size)	-0.18099	0.92916	-2.02692	1.59450
	φ_{13} (Persistence)	-0.35372	0.90625	-2.14058	1.40667
	φ_{20} (Intercept)	-1.09469	0.51498	-2.10679	-0.05940
	φ_{21} (Sex)	-0.61495	0.50759	-1.52946	0.45569
	φ_{22} (Size)	-0.25864	0.63970	-1.53312	0.99333
	φ_{23} (Persistence)	-0.53058	0.87497	-2.28888	1.16768

Assuming the dependence bivariate structure for the Tobit-Weibull model all covariates do not indicate significative effects on both responses of interest (all 95%

credibility intervals include the zero value).

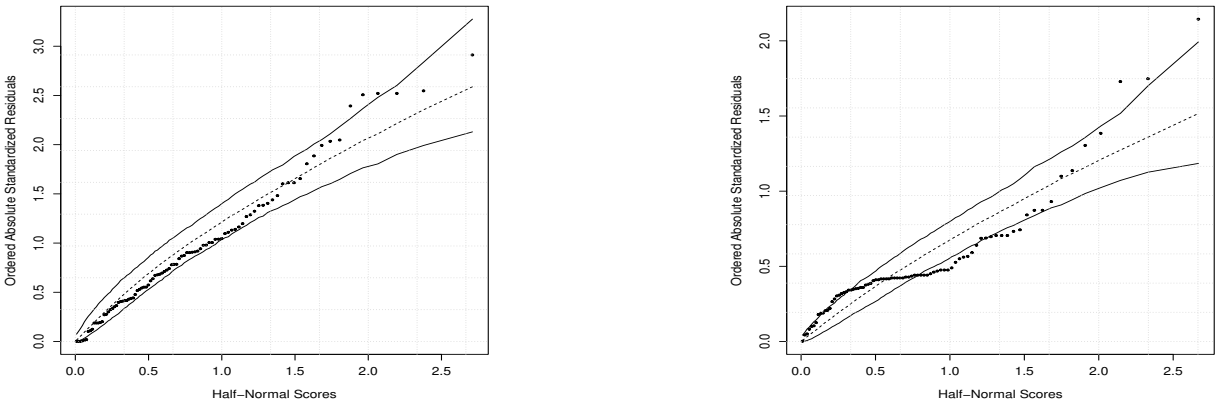

Figure 6 – Envelope for the residuals the Tobit-Weibull model for T1 (left) and T3 (rigth) – thyroid cancer data.

Figure 6 shows the Half-Normal plot with simulated envelope of the residual for the Tobit-Weibull model, assuming the presence of all covariates, for thyroglobulin level before starting therapy ($T1$) and after the last therapy session ($T2$). It is important to point out that for the response thyroglobulim level ($T2$) there are the presence of 62.64% left censored data, which could imply that the proposed model could not be able to detect significative risk factors. as observed in the obtained results of Table 10. However, around of 70% of the observed values are inside of the 95% credible interval for the proposed regression model which is a indication of accuracy even using non-informative *prior* distributions. Moreover, the results could be more accurate assuming informative *prior* distributions, using the reduced model eliminating the non significative factors or removing the outlier's values.

4.3.2 Vaccine Data

In this application, we consider a safety and immunogenicity study of measles vaccines conducted in Haiti during the years 1987-1990 [Job et al., 1991]. The goal of the study was to show that the higher titer vaccines could effectively immunize infants as young as 6 months of age. The immunogenicity analyses indicated much higher antibody responses among high titer recipients, as had been anticipated, and among recipients of the Edmonston-Zagreb vaccine strain as compared to Schwarz strain. Re-analysis of these data was prompted by findings in several countries of higher than expected mortality 2-3 years post vaccination among high titer vaccine recipients, with most of the excess mortality among girls. Neutralization antibody assays were performed on serum from 330 children at 12 months of age [Moulton and Halsey, 1995]. The detection limit was 0.1 international units.

Considering this left-censored data related to vaccine, we fitted the univariate Tobit-Weibull model introduced in Section 4.2. We considered as covariates the type of vaccine used (Schwartz or Edmonston-Zagreb), the level of the dosage (medium or high) and the children's gender (male or female).

The statistical analysis was carried out in the `R` software [R Core Team, 2015] and the `R2jags` package was used to obtain the Bayesian estimates for the model parameters. In this application, we also assume non-informative distributions for the parameters of the proposed model, since we do not have expert opinion. The non-informative distributions assumed in this study are the same as those considered in the previous application. The results for each fitted model are presented in Table 11.

Table 11 – Summary of the fitted Tobit-Weibull model for vaccine data.

Parameter	Estimate	Std. Dev.	95% Cred. Int.	
			Lower	Upper
α	0.6427	0.0544	0.5362	0.7496
γ_0	0.0007	0.2364	-0.4897	0.4382
γ_1 (type)	-0.2285	0.1987	-0.6208	0.1608
γ_2 (level)	-0.2942	0.1973	-0.6804	0.0948
γ_3 (sex)	0.2597	0.1906	-0.1116	0.6347
φ_0	-0.5991	0.2203	-1.0365	-0.1724
φ_1 (type)	-0.7181	0.2549	-1.2246	-0.2251
φ_2 (level)	-0.3877	0.2481	-0.8748	0.0956
φ_3 (sex)	0.0810	0.2302	-0.3760	0.5290

Based on the obtained results assuming the Bayesian estimates in Table 11, we observe from the obtained 95% credible intervals (zero value is not inside the intervals), that the neutralization antibody levels are affected by type of vaccine used (Schwartz or Edmonston-Zagreb) assuming the logistic structure.

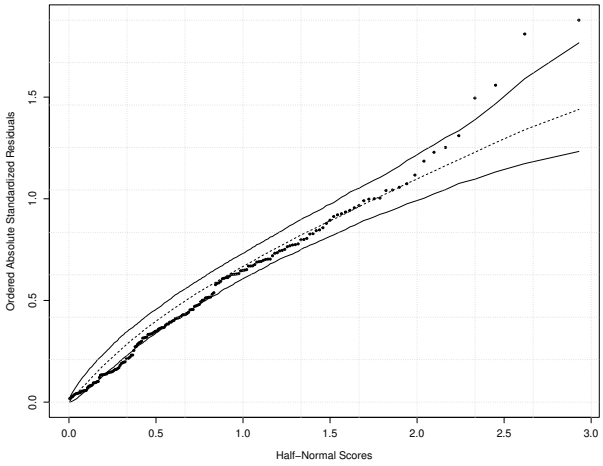

Figure 7 – Envelope for the residuals the Tobit-Weibull model for vaccine data.

Figure 7 shows the Half-Normal plot with simulated envelope of the residual for the Tobit-Weibull model, assuming the presence of all covariates. We observed that despite the presence of about 57.6% left censored observations, the predictions obtained for both models were well adjusted. Most of these values are within the 95% credibility range for both the model without covariate and the proposed regression model, which is an indication of good accuracy.

4.4 Concluding remarks

This study introduced univariate and bivariate models based on the Tobit-Weibull distribution as an alternative to the left-censoring data analysis. The proposed model was considered to analyze two medical survival dataset related to cancer and vaccine. We found promising results in discovering the significant risk factors affecting the survival times.

We have considered a fully Bayesian approach for model estimation. The adopted method was based on the Metropolis-within-Gibbs algorithm for sampling pseudo-random values from the *posterior* distribution of model parameters. We acknowledge the remarkable computational simplicity of estimating cure rate models baseline by the Tobit-Weibull distribution. Furthermore, Bayesian techniques allow incorporating *prior* information from experts, leading to much more insightful inferential results. Besides, such methods also provide straightforward interpretations based on the estimated model parameters, which is a prominent concern in medical applications.

Using the proposed models, we found the results obtained from univariate analysis

is more accurate than the multivariate. This fact is directly related to adopting a regression structure or the dependence structure between both times. Our study identified only two significant factors for the analyzed data. Noticeably, the observed differences highlight the importance of using appropriate statistical models, especially in clinical studies. These models can be an alternative to the widely used Cox proportional hazards model.

Chapter 5

Use of a Tobit-Weibull model in the analysis of daily rain precipitation data for São Paulo city, Brazil (2007- 2021)

5.1 Introduction

The climatic changes observed in recent decades have led to great concern, as the effects of these changes could be catastrophic across the planet. From different statistical analyzes for climatic data collected in climate stations around the planet, it is possible to observe the loss of sea ice, accelerated sea level rise and longer and more intense heat waves around the world, as pointed out by the Intergovernmental Panel on Climate Change [IPCC, 2007, 2013].

The effects of climate change has being observed worldwide particularly in temperature and rain precipitation. Many papers were introduced in the literature in recent decades related to climate change (precipitation, temperature, level of the oceans among many others) and its implications (see for example, Arnell, 2014, Alexander et al., 2006, Bonan, 2008, Costello et al., 2009, Hawkins et al., 2017, Lineman et al., 2015, Kabir et al., 2016, Kaczan and Orgill Meyer, 2020, Levermann et al., 2013, Zhiying and Fang, 2016, Matthews, 2018, Poloczanska et al., 2013, Rahmstorf et al., 2007, Serdeczny et al., 2017, Springmann et al., 2016, Turner et al., 2020, Karl, T. R. et al., 2009, Zhao et al., 2017, Richards, 1993).

Climatic variables, in particular, the occurrence of rainfall and its intensity, have a great impact on populations, especially in agriculture and urban centers where great

irregularity in rainfall has been observed in the last years, sometimes with long dry periods and lack of water for agriculture and urban centers and other times with long wet periods with large amounts of rain in small areas leading to great floods, great destruction and great human losses. The statistical modeling of precipitation data (occurrence and intensity) is needed for forecasting, planning, and also for the management of water resources systems [Dzupire et al., 2018]. In addition, the amount of rainfall and its occurrence is fundamental for agricultural production [Lobell and Burke, 2010].

Rainfall data in general are given by binary data (occurrence or not of precipitation in a day) and a positive real measure (total amount of rain in a given day). In practice, statistical modeling of rain precipitation has an important particularity: the occurrence of excess of zeros, that is, the occurrence of many dry days (no rain) that can make it difficult to use traditional statistical techniques, possibly using parametrical models in presence of standard left-censoring mecanisms. This requires the use of appropriate probability models to describe precipitation data. Some work linked to daily precipitation data can be viewed at Wilks, 1998, Dzupire et al., 2018, Benestad et al., 2019, Auestad et al., 2012, Stern and Coe, 1984, Yeo et al., 2019, George et al., 2016, Bárdossy et al., 2021, Latifoglu, 2021.

In this Chapter, the response variable we will work with is the total daily precipitation collected at a climate station located in the city of São Paulo. A Tobit-Weibull model is fitted, under a Bayesian approach, to verify if the behavior of the total daily rainfall in the period, are changing in the follow-up period, that is, if there are linear and quadratic effects in the covariates years and months. A logistic regression model for the occurrence (or not) of daily rainfall will be fitted in the mixture component of the Tobit-Weibull model. Other climate variables such as daily mean atmospheric pressure, daily mean temperature and daily mean humidity are also analyzed considering standard regression models with normal errors or Weibull distributions for the case of asymmetric data as observed for daily mean humidity. Thus, the Section 5.2 introduces the dataset; Section 5.2 presents the Tobit model assuming a Weibull distribution; Section 5.4 presents the obtained results; finally Section 5.5 presents some concluding remarks.

5.2 Materials and Methods

Tobit model assuming a Weibull distribution

Tobin [1958] proposed a methodology, named the Tobit model, that could fit the data appropriately by assuming a regression model whose response variable was censored

to a pre-fixed limiting value or could be given by a repeated fixed number. This is the case with daily rain precipitation data where there are many days in a month (or year) without or with very small amount of rain precipitation, that is, with zero values in the dataset. The censoring occurs when the response of the regression model is not directly observable, but its independent variables (or covariates) are observed.

Tobit models usually rely on the normality assumption. Proposals of Tobit models that relax this assumption are extremely important, since it is common knowledge that most of the data available in the real world are often well modeled by non-normal distributions. A number of authors have noticed that the asymmetry of data of censored responses and their kurtosis usually are different from the expected for a Normal distribution, so that more flexible Tobit models are needed (see, for example, Martínez-Flórez et al. [2013]).

From the definition of the indicator of censoring variable, $\delta = 1$ if T (total precipitation observed in a day) is a complete observation $(T > C)$ and $\delta = 0$ if T is a left censored observation $(T \leq C)$. Here we assume as left-censoring the zero value (no precipitation in a day) and C is arbitrary fixed as the value 0.01.

If we have a complete observation, that is, $(T > C)$, in this work we assume a truncated generalized form of the Weibull distribution for the random variable T with probability density function given by,

$$f(t \mid T > C) = \frac{f_0(t)}{P(T > C)} \tag{5.1}$$

where

$$f_0(t) = \frac{\alpha}{\beta^\alpha} t^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\}$$

and, the probability of $T > t$ is given by,

$$S_0(t) = P(T > t) = \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\}$$

Remark: $S_0(C) = P(T > C) = \exp\left\{-\left(\frac{C}{\beta}\right)^\alpha\right\}$ where C is a known constant.

Let us assume a mixture model, given by the probability density function,

$$f(t) = p[dC(t)] + (1-p)\frac{f_0(t)}{S_0(C)} \tag{5.2}$$

where $f_0(t) = \frac{\alpha}{\beta^\alpha} t^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\}$ and $S_0(C) = \exp\left\{-\left(\frac{C}{\beta}\right)^\alpha\right\}$ and $dC(t)$ is the Dirac measure at $(0, C)$ (in the mixture model, we are assuming a degenerated probability distribution $P(T = 0) = 1$).

Remarks: $S_0(C) = P(T > C) = \exp\left\{-\left(\frac{C}{\beta}\right)^\alpha\right\}$ where C is a known constant.

- If $T \leq C$, $S(t) = P(T > t) = p + (1-p)\int_C^\infty \frac{f_0(u)du}{S_0(C)} = p + (1-p)\frac{S_0(C)}{S_0(C)} = p + (1-p) = 1$, where $p = P(T \leq C)$ and $1 - p = P(T > C)$
- If $T > C$, $S(t) = P(T > t) = 1 - P(T \leq C) = 1 - \{P(0 < T < C) + P(C < T < t)\} = 1 - \{p + (1-p)\int_C^t \frac{f_0(u)du}{S_0(C)\}}$

where $\int_C^t f_0(u)du = P(T > C) - P(T > t) = S_0(C) - S_0(t)$. Thus,

$$S(t) = 1 - \{p + \frac{(1-p)}{S_0(C)}[S_0(C) - S_0(t)]\} = 1 - \{p + (1-p) - [(1-p)\frac{S_0(t)}{S_0(C)}]\}$$

That is,

$$S(t) = (1-p)\frac{S_0(t)}{S_0(C)}$$

In summary, the survival function is given by,

- If $T \leq C$, $S(t) = 1$
- If $T > C$, $S(t) = (1-p)\frac{S_0(t)}{S_0(C)}$

where $S_0(t) = \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\}$ and $S_0(C) = \exp\left\{-\left(\frac{C}{\beta}\right)^\alpha\right\}$.

The likelihood function for the parameters p, α and β based on one observation t is given from (5.2), by,

$$L(p, \alpha, \beta) = p[dC(t)] + (1-p)\frac{f_0(t)}{S_0(C)} \tag{5.3}$$

where $f_0(t) = \frac{\alpha}{\beta^\alpha} t^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\}$ and $S_0(C) = \exp\left\{-\left(\frac{C}{\beta}\right)^\alpha\right\}$.

With the censoring information, let us define a binary variable $\delta = 1$ if T is a complete observation $(T > C)$ and $\delta = 0$ if T is a left censored observation $(T \leq C)$ with conditional probabilities given by,

$$\begin{aligned} P(\delta = 0 \mid p, \alpha, \beta, t) &= \frac{p}{p+(1-p)\frac{f_0(t)}{S_0(C)}} \\ \text{and,} & \\ P(\delta = 1 \mid p, \alpha, \beta, t) &= \frac{(1-p)\frac{f_0(t)}{S_0(C)}}{p+(1-p)\frac{f_0(t)}{S_0(C)}} \end{aligned} \tag{5.4}$$

In this way, we have a Bernoulli distribution with probability function,

$$P(\delta) = \left\{\frac{p}{p + (1-p)\frac{f_0(t)}{S_0(C)}}\right\}^{(1-\delta)} \left\{\frac{(1-p)\frac{f_0(t)}{S_0(C)}}{p + (1-p)\frac{f_0(t)}{S_0(C)}}\right\}^{\delta} \tag{5.5}$$

where $\delta = 1 \quad (T > C) \quad$ or $\quad \delta = 0 \quad (T \leq C)$.

The likelihood function $L(p, \alpha, \beta)$ based on n observations is given by,

$$L(p, \alpha, \beta; \mathbf{t}) = \prod_{i=1}^{n} p^{(1-\delta_i)} \left[(1-p)\frac{f_0(t_i)}{S_0(C)}\right]^{\delta_i} \tag{5.6}$$

Assuming the truncated Weibull distribution, the log-likelihood function for p, α and β is given (from (5.6)) by,

$$\begin{aligned} \ell(p, \alpha, \beta; \mathbf{t}) &= \sum_{i=1}^{n} \Bigg\{ (1-\delta_i)\log(p) + \delta_i \Bigg\{ \log(1-p) + \log(\alpha) + \\ &+ (\alpha - 1)\log(t_i) - \alpha\log(\beta) - \left(\frac{t_i}{\beta}\right)^\alpha + \left(\frac{C}{\beta}\right)^\alpha \Bigg\} \Bigg\} \end{aligned} \tag{5.7}$$

Remark: For the daily precipitation data we have $\delta = 1$ (day with rain precipitation) or $\delta = 0$ (day without rain precipitation, that is, $T = 0$).

Inferences for the parameters of the models are obtained under a Bayesian approach using existing MCMC methods like the as Metropolis-within-Gibbs algorithms. In this way, in the simulation of samples of the joint *posterior* distribution [Gelfand and Smith, 1990], $\pi(\theta|data)$ where θ is the vector of all parameters, we use Gibbs or Metropolis-Hastings

algorithms, where it is needed to sample each parameter from the *posterior* conditional distributions $\pi(\theta_r|\theta_{(r)})$, where $\theta_{(r)}$ denotes the vector of all parameters except θ_r and r is associated to each one of the parameters of the model. To simplify the computational work in the iterative procedure to get the Bayesian inferences, the literature presents different free softwares to simulate samples of the joint *posterior* distribution of interest.

For a Bayesian analysis it as assumed uniform $U(a, b)$ *prior* probability distributions for the parameters α and β with a and b known hyperparameters and a $Beta(e, f)$ or a $U(0, 1)$ *prior* probability distribution for p with e and f known hyperparameters. We further assume *prior* independence among the three parameters.

In presence of a vector of covariates $x = (x_1, x_2, \ldots, x_p)^\top$ let us assume a regression model for the scale parameter *beta* given by,

$$\beta = \exp(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \ldots + \gamma_p x_p) \tag{5.8}$$

where $\gamma = (\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_p)^\top$ is the regression parameter vector associated to covariate vector $\mathbf{x} = (x_0, x_1, x_2, \ldots, x_p)^\top$ and a logit model for the parameter p, given by,

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \varphi_0 + \varphi_1 x_1 + \varphi_2 x_2 + \ldots + \varphi_p x_p \tag{5.9}$$

For a Bayesian analysis it as assumed a uniform $U(a, b)$ *prior* distribution for the parameter α, Normal $N(c, d^2)$ *prior* distributions for the regression parameters $\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_p$ with c and d known hyperparameters and Normal $N(e, f^2)$ *prior* distributions for the regression parameters $\varphi_0, \varphi_1, \varphi_2, \ldots, \varphi_p$ with e and f known hyperparameters.

For a Bayesian approch of the proposed models we use MCMC simulation methods to get the *posterior* summaries of interest (see, for example, Chib and Greenberg, 1995, Gelfand and Smith, 1990, Gelman et al., 1995, Geman and Geman, 1984, Gilks et al., 1995).

5.3 Application to a climatic dataset

The original climatic data considered in this study (total rainfall per hour in mm; atmospheric pressure at the station level in milibar (mB); temperature per hour in ^oC; relative air humidity per hour in %) was obtained from the Mirante climate station located in the city of São Paulo, Brazil (latitude = -23.59; longitude = -46.52 and altitude = 785.64 meters) for the period from January 1, 2007 to December 31, 2021 (n = 5479

measurements). Dataset obtained in https://portal.inmet.gov.br/dadoshistoricos. This site presents climate data for several cities in Brazil. Missing observations were imputed as means (previous and *posterior* observed values) for the variables temperature and humidity; for total missing precipitation, the value zero was considered. The possible limitation of the data is related to an apparently short period (15 years) as many weather stations in Brazil given by INMET (Instituto Nacional de Meteorologia, Brasil) were created recently, a common fact in third world countries where it is usually difficult to obtain longer series of climatic observations.

Table 12 shows the averages of the hourly climate variables per year, standard deviations, medians, maximums and minimums in each year (2007 to 2021). The number of observations per year is given by N. Figure 8 shows the boxplot of the observed values in each year for each climate variable.

Table 12 – Hourly means, medians, standard-deviations, maximuns and minimuns in each year (2007 to 2021)

Climate variable	Year	N	Mean	Std. Dev.	Minimum	Median	Maximum
Total rainfall	2007	8760	0.1757	1.4275	0.0000	0.0000	54.4000
	2008	8784	0.1777	1.4660	0.0000	0.0000	54.4000
	2009	8760	0.2242	1.5771	0.0000	0.0000	42.4000
	2010	8760	0.2093	1.6064	0.0000	0.0000	57.6000
	2011	8760	0.1887	1.5852	0.0000	0.0000	58.4000
	2012	8784	0.2082	1.5330	0.0000	0.0000	45.6000
	2013	8760	0.1552	1.0124	0.0000	0.0000	29.6000
	2014	8760	0.1393	1.2685	0.0000	0.0000	45.2000
	2015	8760	0.2158	1.7504	0.0000	0.0000	77.8000
	2016	8784	0.1724	1.3696	0.0000	0.0000	39.6000
	2017	8760	0.1849	1.5494	0.0000	0.0000	64.6000
	2018	8760	0.1348	1.0894	0.0000	0.0000	29.2000
	2019	8760	0.2000	1.5403	0.0000	0.0000	67.6000
	2020	8782	0.1971	1.5171	0.0000	0.0000	44.4000
	2021	8759	0.1376	1.0254	0.0000	0.0000	31.0000
Atmospheric pressure	2007	8760	926.84	3.64	915.60	926.60	938.50
	2008	8784	926.51	3.55	915.10	926.50	938.00
	2009	8760	926.27	3.31	915.10	926.10	935.60
	2010	8760	926.74	3.98	913.10	926.40	938.60
	2011	8760	926.48	3.67	915.50	926.50	938.40
	2012	8784	927.11	3.68	916.90	926.70	938.10
	2013	8760	927.05	3.42	917.80	926.90	937.50
	2014	8760	927.53	3.27	918.30	927.10	938.60
	2015	8760	927.34	3.32	916.30	927.20	940.30
	2016	8784	927.35	3.65	914.90	927.20	938.60
	2017	8760	927.60	4.03	915.30	927.40	940.30
	2018	8760	927.18	3.58	916.60	927.00	937.30
	2019	8760	927.48	3.52	918.30	927.10	939.40
	2020	8783	927.23	3.51	914.70	927.00	938.40
	2021	8759	926.92	3.71	915.70	926.70	938.90
Temperature	2007	8760	20.340	4.660	5.200	20.300	34.300
	2008	8784	19.607	4.171	8.500	19.300	33.800
	2009	8760	20.210	4.270	6.800	20.100	33.400
	2010	8760	20.128	4.578	8.800	20.100	33.300
	2011	8760	19.847	4.636	6.500	19.700	33.300
	2012	8784	20.509	4.399	8.900	20.100	36.000
	2013	8760	19.927	4.426	5.700	19.700	33.900
	2014	8760	21.060	4.704	9.200	20.600	37.000
	2015	8760	21.078	4.312	11.100	20.700	36.300
	2016	8784	20.336	4.831	4.100	20.300	35.200
	2017	8760	20.455	4.351	7.900	20.300	34.600
	2018	8760	20.454	4.286	9.200	20.200	33.200
	2019	8760	21.045	4.651	6.700	20.800	35.700
	2020	8783	20.474	4.461	8.300	20.100	37.300
	2021	8759	20.083	4.603	4.400	19.700	35.400
Relative air humidity	2007	8760	70.780	17.416	16.000	76.000	97.000
	2008	8784	71.690	16.496	14.000	77.000	96.000
	2009	8760	73.822	15.319	10.000	79.000	95.000
	2010	8760	71.536	18.878	13.000	78.000	97.000
	2011	8760	73.564	19.245	12.000	80.000	100.000
	2012	8784	72.225	18.634	12.000	78.000	100.000
	2013	8760	73.172	18.296	15.000	79.000	98.000
	2014	8760	67.290	18.936	13.000	72.000	97.000
	2015	8760	69.653	16.634	12.000	75.000	95.000
	2016	8784	68.467	16.226	13.000	74.000	95.000
	2017	8760	68.059	16.661	12.000	73.000	94.000
	2018	8760	68.504	15.888	15.000	74.000	92.000
	2019	8760	67.861	16.666	14.000	73.000	92.000
	2020	8783	67.484	17.652	13.000	73.000	96.000
	2021	8758	68.345	17.876	12.000	74.000	98.000

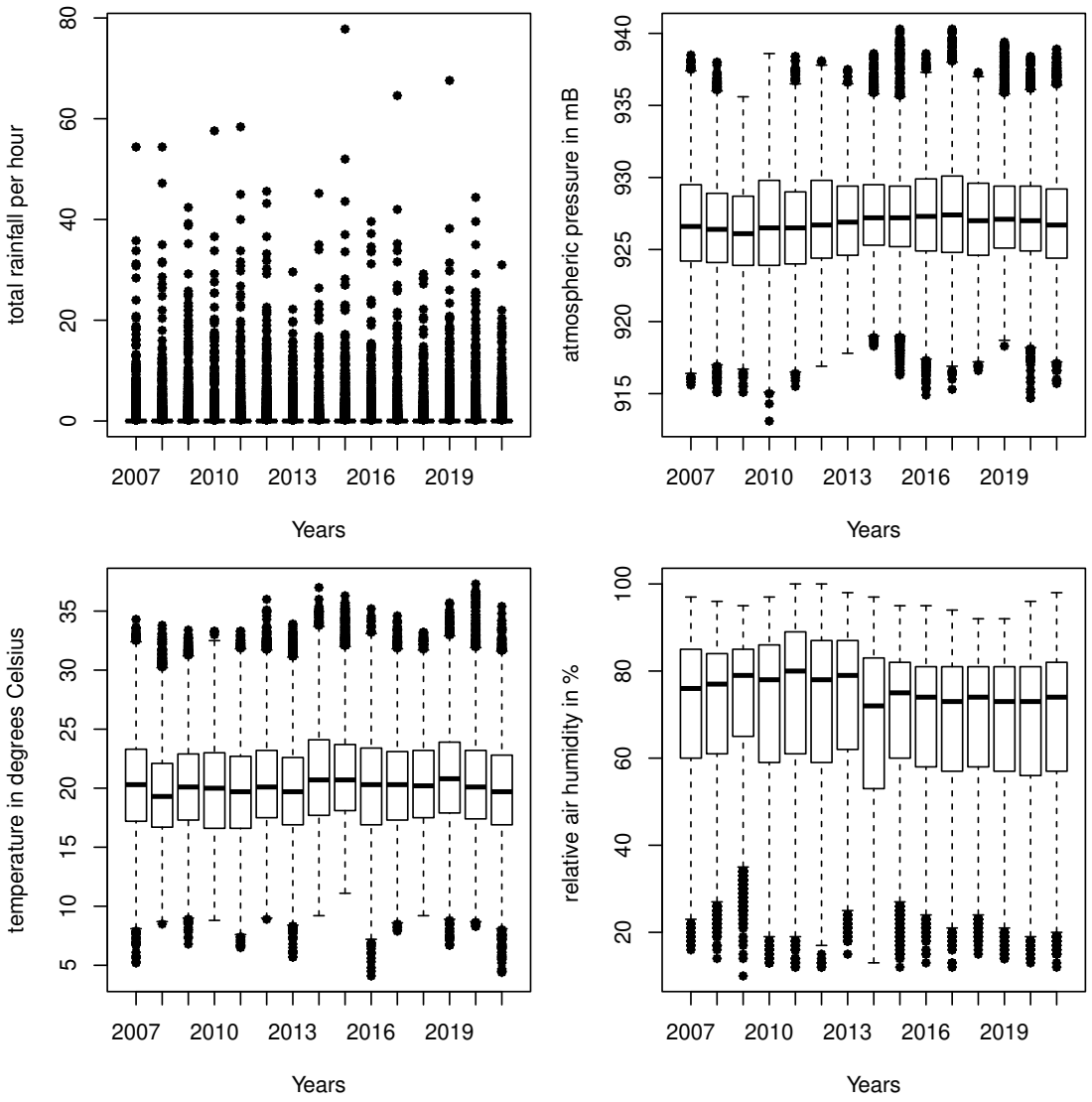

Figure 8 – Maximum measures in each year for climate variables obtained per hour (2007 to 2021).

In general, the greatest implications of climate change are caused by the maximums of climatic variables observed in each hour of a large period of time, such as the maximum values of precipitation per hour or the maximum values of temperature per hour. Figure 10 shows the graphs of the maximum observed in each year for each climate variable (the plots also have fitted polynomial models of order 3).

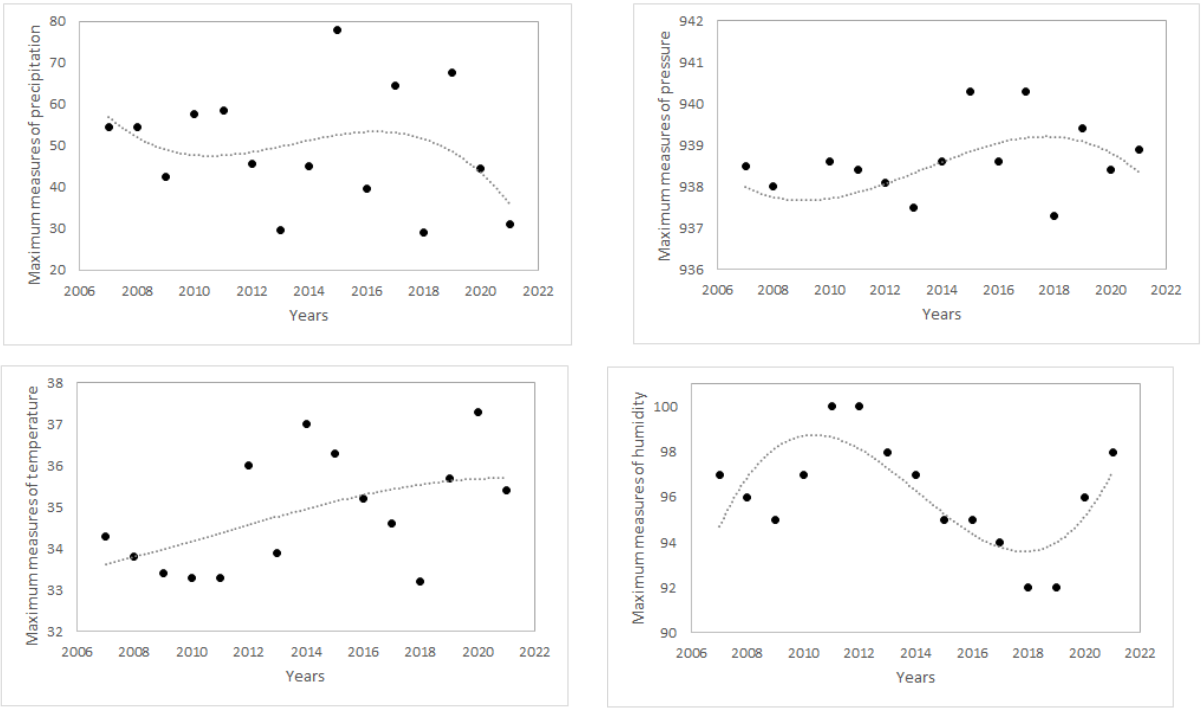

Figure 9 – Maximum measures in each year for climate variables obtained per hour (2007 to 2021).

From the graphs in Figure 9, we have some preliminary conclusions:

- A very large maximum hourly precipitation value was observed in the year 2015 ($77.8000mm/h$), which is a very large volume of rain in a very short period of time that can have catastrophic effects for the locality.
- A maximum value of atmospheric air pressure per hour was also observed in the years 2015 and 2017 ($940.3mB/h$).
- The maximum hourly temperature value has been increasing consistently in the period from 2007 to 2021, an indication that the temperature of the city of São Paulo has been increasing in recent years despite the short follow-up period.
- The maximum values of humidity per hour increased until the year 2011; after that year there was a fall; in the last years of the observed period there is an increase in the maximum humidity per hour.

Table 13 shows the total rainfall for each year (sum of the amount of rain observed in each hour) and the total number of hours in each year with the presence of rain in the

period from 2007 to 2021, from which it is observed that in the year 2015, we have the highest amount of total rainfall (1890.6 mm) observed in the 15-year follow-up period. Likewise, it is observed that the number of hours with the presence of rain has become smaller in the last years of follow-up.

Table 13 – Rainfall totals for each year (sum of the amount of rain observed in each hour) and the total number of hours with the presence of rain in each year in the period from 2007 to 2021.

Years	Hours	Total rainfall	Hours with rainfall
2007	8760	1539.4	663
2008	8784	1561.0	668
2009	8760	1964.4	856
2010	8760	1833.6	794
2011	8760	1653.2	661
2012	8784	1828.8	782
2013	8760	1359.6	738
2014	8760	1220.6	528
2015	8760	1890.6	753
2016	8784	1514.4	658
2017	8760	1619.6	671
2018	8760	1180.8	612
2019	8760	1752.2	679
2020	8782	1730.8	677
2021	8759	1205.2	612

Figure 10 shows the graphs of the total hours with the presence of rain in each year in the period from 2007 to 2021, where a decline in the number of hours in the year with the presence of rain can be observed. We also have in Figure 10, the graph of the total precipitation accumulated in each year in the period from 2007 to 2021.

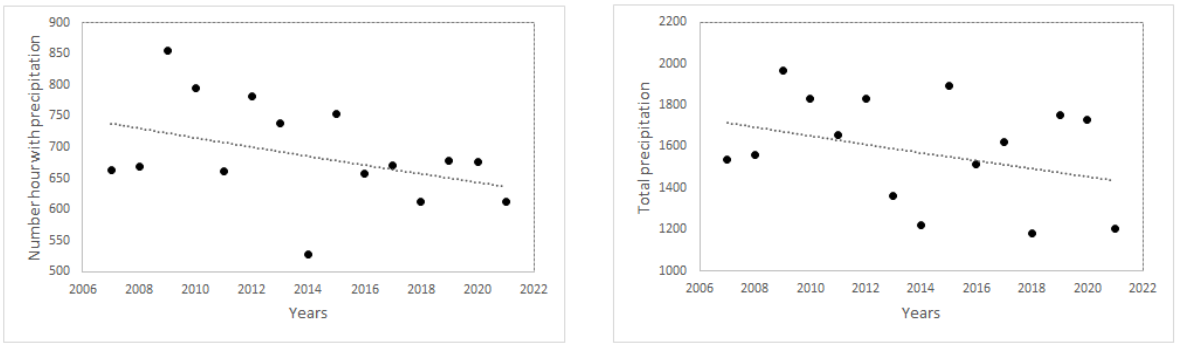

Figure 10 – Total hours with presence of rain in each year and total precipitation accumulated in each year in the period from 2007 to 2021.

Daily climatic data

From the hour climate data of São Paulo city, we obtained the daily climate data for the period $2007 - 2021$. Histograms of daily precipitation, daily mean temperature, daily mean air pressure and daily mean humidity in the period from 2007 to 2021 are presented in Figure 11. From these histograms, we observe that the histogram of the daily precipitations shows a zero excess indicating that it is needed an appropriate model which captures this particularity (2323 days with rain and 3156 days without rain; total of $n = 5479$ days). It is important to point out that we could assume in a first modeling approach, the zero data (days with no rain) as left censored data with the same probability distribution as the complete data, but this approach could lead to very poor inferences when there is a great proportion of left censored data as in our case of daily precipitation data.

For the cases of daily mean air pressure and daily mean temperatures we see from Figure 11, approximately symmetry for the histograms, indicating the possibility to fit a Normal distribution for the data. For the case of daily mean humidity we see the need to fit an asymmetrical model, as for example, a Weibull distribution with two parameters. Figure 12 shows the scatter plots of daily rain precipitation, daily mean temperature, daily mean air pressure and daily mean humidity versus months in the period from 2007 to 2021, from where it is observed the need of a statistical model wich captures the quadratic effects of months in the climate variables.

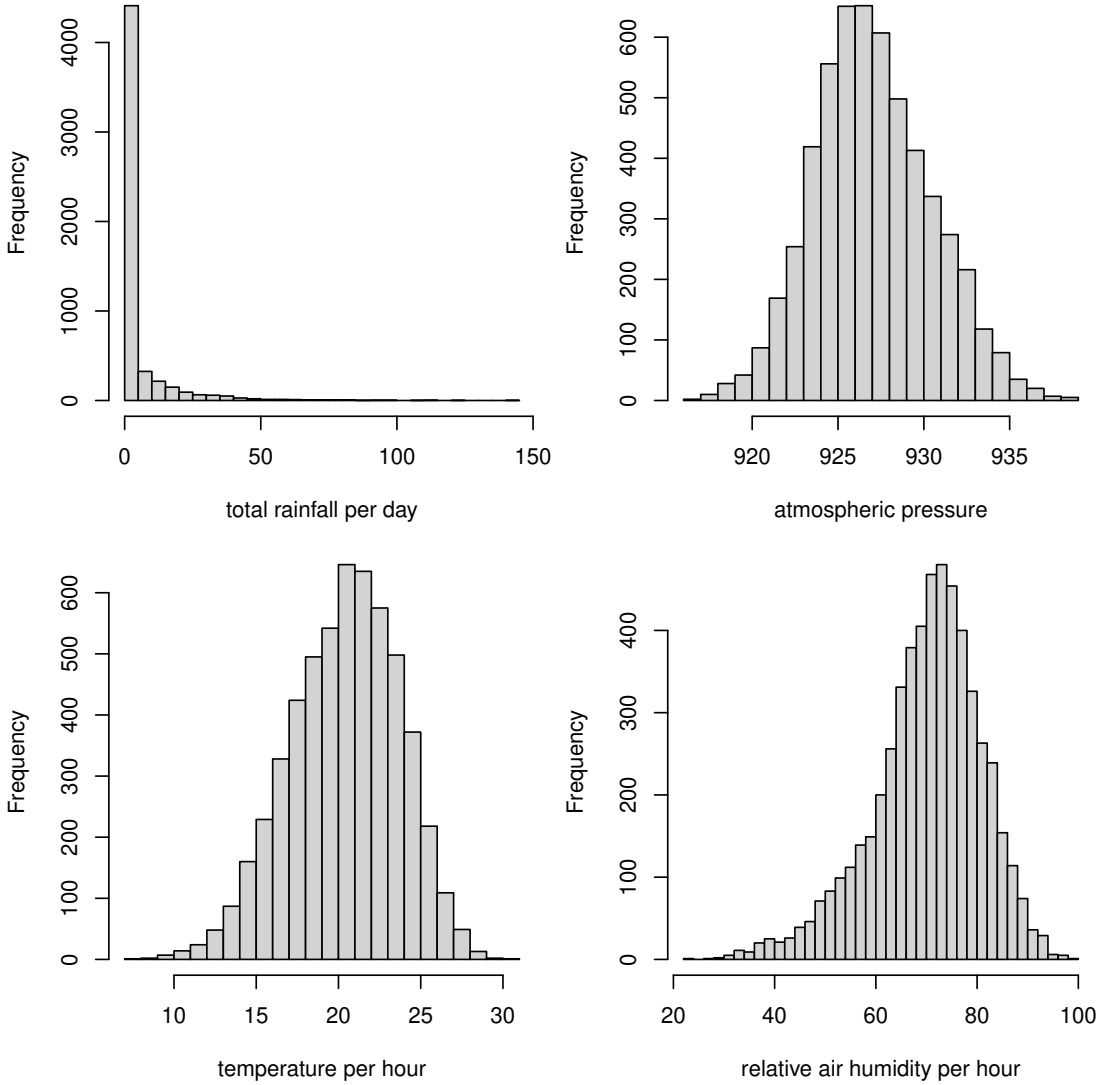

Figure 11 – Histograms of daily precipitation, daily mean temperature, daily mean air pressure and daily mean humidity in the period from 2007 to 2021.

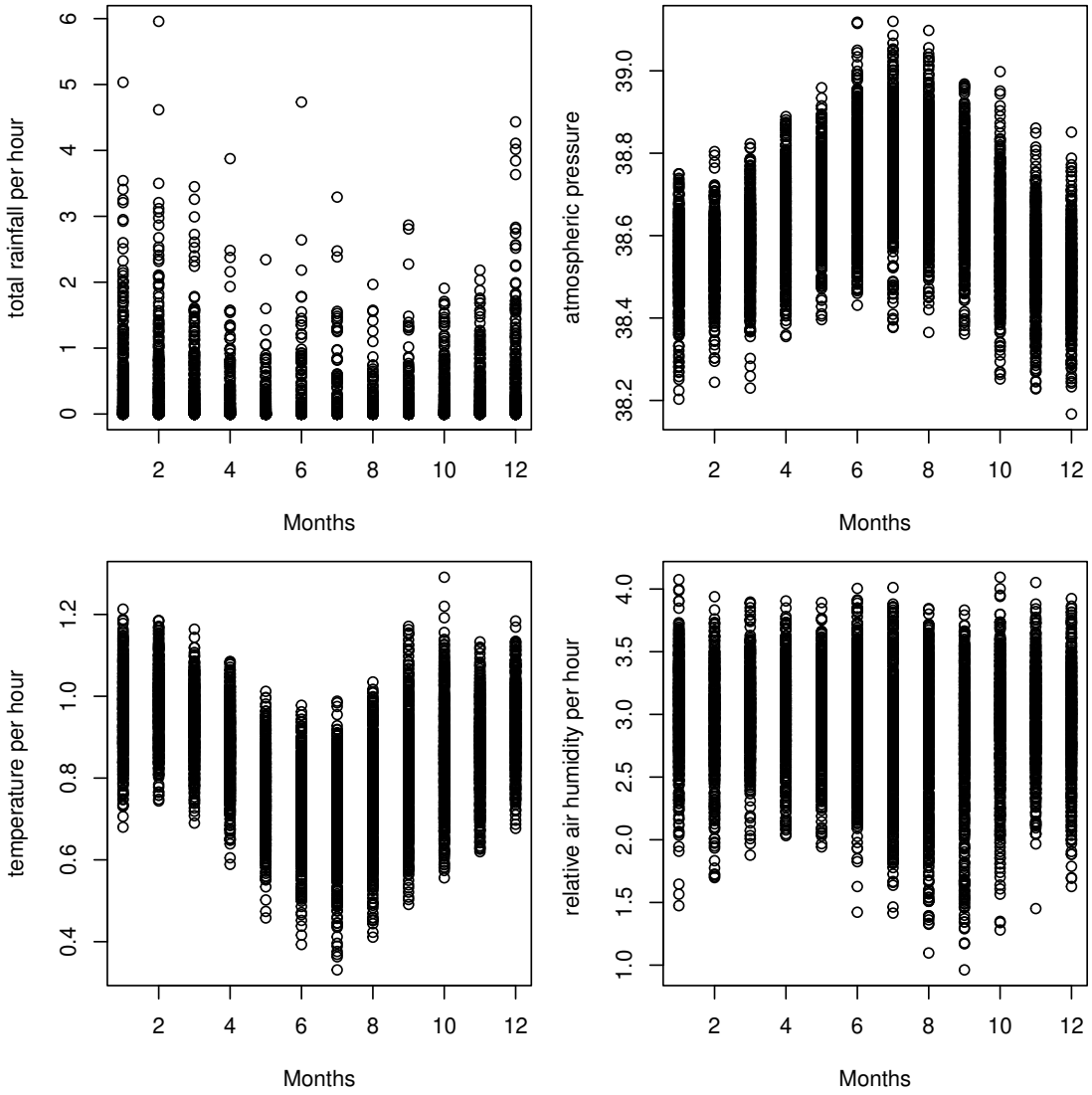

Figure 12 – Scatter plots of daily precipitation, daily mean temperature, daily mean air pressure and daily mean humidity versus months in the period from 2007 to 2021.

5.4 Results

Daily rain precipitation

Assuming the Tobit-Weibull model defined by (5.2), for the daily precipitation data not considering the presence of covariates we considered the following prior distributions for the parameters α, β and p: $\alpha \sim U(0.1, 2)$, $\beta \sim U(0.1, 100)$ and $p \sim U(0, 1)$. We also assumed *prior* independence among the parameters. Using the `R` software [R Core Team, 2015] and the `R2jags` package, we simulated a total of $11,000$ Gibbs samples (the first

$1,000$ samples considered as a burn-in-sample deleted to eliminate the effect of the initial values and using $1,000$ final samples chosen from every 10^{th} sample) to get the *posterior* summaries of interest from the joint *posterior* distribution for α, β and p. Convergence of the simulation algorithm was verified from trace plots of the generated samples for each parameter. The computer code is presented in Appendix 7. Table 14 shows the *posterior* summaries of interest. Considering the data with observed rain precipitations the mean and median of the Weibull distribution (5.1) are given, respectively, by, $E(T) = \beta\Gamma(1 + 1/\alpha)$ and $median = \beta(log(2))^1/\alpha$ (obtained from the equation $S(t) = \frac{1}{2}$). Thus, from the Bayes estimators we get the Bayesian estimatives for the mean given by 10.1842 and for the median given by 4.05991. The sample mean and sample median are given, respectively, by 10.269 and 4.000. That is, we have an indication of good fit for the dataset. Also the observed proportion of days with rain is given by $2323/5479 = 0.42398$ (close to the Bayesian estimator of $1 - p$ given by 0.4240).

Figure 13 shows the histogram of the data associated to days with rain and the fitted Weibull density with parameters $\alpha = 0.6628$ and $\beta = 7.6353$. From this figure we observe a very good fit of the Weibull distribution to the dataset (total rain precipitation in each day).

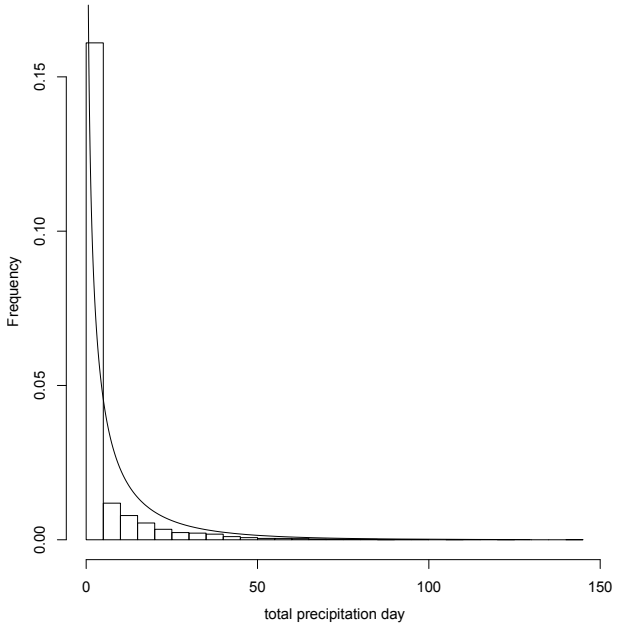

Figure 13 – Histogram of the data (total daily precipitation) and the fitted Weibull density with parameters $\alpha = 0.6628$ and $\beta = 7.6353$

Now, assuming the Tobit-Weibull model defined by (5.2) and (5.8), for the daily precipitation data in presence of covariates years (linear effects) and months (linear and

quadratic effects) we considered the following regression models:

$$\beta_i = \gamma_0 + \gamma_1(years_i) + \gamma_2(months_i) + \gamma_3(months_i)^2 \tag{5.10}$$

and

$$logit(p_i) = \zeta_0 + \zeta_1(years_i) + \zeta_2(months_i) + \zeta_3(months_i)^2 \tag{5.11}$$

where $i = 1, 2, \ldots, 5479$ (number of days from January 01, 2007 to 31 December, 2021).

Table 14 – *Posterior* summaries for the Tobit-Weibull model (daily rain precipitation data)

Parameter	Estimate	Std. Dev.	95% Cred. Int.	
			Lower	Upper
α	0.6506	0.0108	0.6294	0.6714
γ_0	7.7432	0.5616	6.6966	8.8188
γ_1	0.0796	0.0486	-0.0145	0.1753
γ_2	-0.6475	0.2036	-1.0697	-0.2542
γ_3	0.0498	0.0161	0.0201	0.0830
ζ_0	-1.7949	0.1124	-2.0082	-1.5689
ζ_1	0.0069	0.0067	-0.0064	0.0201
ζ_2	0.7940	0.0360	0.7207	0.8584
ζ_3	-0.0572	0.0027	-0.0619	-0.0518

We assumed the following *prior* distributions for the parameters α, γ_j and ζ_j: $\alpha \sim U(0.1, 2)$, $\gamma_0 \sim N(0, 1)$, $\zeta_0 \sim N(0, 1)$, $\gamma_j \sim N(0, 10)$,and $\zeta_j \sim N(0, 10)$, $j = 1, 2, 3$. We also assumed *prior* independence among the parameters. Using the `R` software ([R Core Team, 2015]), we first simulated a total of $121,000$ Gibbs samples (the first $101,000$ samples considered as a burn-in-sample deleted to eliminate the effect of the initial values and using $1,000$ final samples chosen from every 20^{th} sample) to get the *posterior* summaries of interest. Convergence of the simulation algorithm was verified from trace plots of the generated samples for each parameter. The computer code is presented in Appendix 7. Table 14 shows the *posterior* summaries of interest. From the results, we observe that the covariate months (linear effects) show significant effects on the scale of the Weibull distribution (negative linear effects on months, that is, decreasing in daily total precipitations in the last months of the year) since the value zero is not inside the 95% credible interval for the regression parameter ζ_2; the same conclusions are observed for the probabilities to have days with rain.

Daily mean air pressure

Considering the daily mean air pressure data in São Paulo city in the same period of time $(2007 - 2021)$ we assumed a Normal distribution $N(\mu, \sigma^2)$ in the data analysis, motivated by the histogram of the daily mean air pressure presented in Figure 14. Under a Bayesian approach we assumed the following *prior* (data not considering the presence of covariates) distributions for the parameters of the model: $\mu \sim N(900, 0.001)$ and $\tau = 1/s = \sigma^2 \sim U(0, 1)$. From the `R` software ([R Core Team, 2015]) (burn-in sample $1,000$ and $1,000$ additional Gibbs samples (every 10^{th} from $10,000$ generated samples) we obtained the Monte Carlo estimates for the *posterior* means of μ and τ (between parentheses the corresponding 95% credible intervals) given, respectively by 927.00 $(926.9; 927.1)$ and 0.0850 $(0.0819; 0.0879)$. That is, the variance σ^2 is estimated by 11.7578 and the standard-deviation σ is estimated by 3.42896. It is interesting to observe that the sample average and sample deviation of the daily mean air pressure data are given, respectively, by 927.04 and 3.43. That is, we have a indication of excelent fit of the Normal distribution for the data. Figure 14 shows the histogram of the daily mean air pressure and the fitted Normal density with parameters $\mu = 927$ and $\sigma = 3.43$. We observe a very good fit of the Normal distribution to the daily mean air pressure.

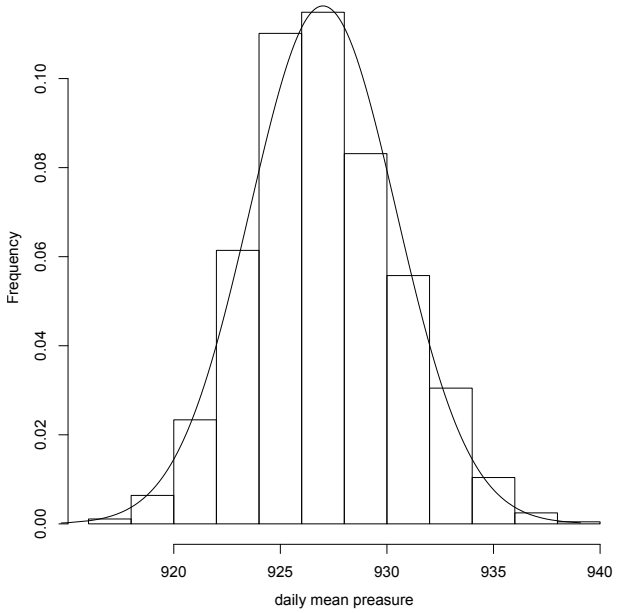

Figure 14 – Histogram of the data (daily mean air pressure) and the fitted Normal density with parameters $\mu = 927$ and $\sigma = 3.43$

Assuming the presence of covariates years (linear effects) and months (linear and quadratic effects) for the daily mean air pressure data we considered the following regression

model in the mean μ of the Normal distribution $N(\mu, \sigma^2)$:

$$\mu_i = \beta_0 + \beta_1(years_i) + \beta_2(months_i) + \beta_3(months_i)^2 \tag{5.12}$$

where $i = 1, 2, \ldots, 5479$ (number of days from January 01, 2007 to 31 December, 2021).

We assumed the following *prior* distributions for the parameters β_0, β_j, $j = 1, 2, 3$ and $\tau = 1/\sigma^2$: $\beta_0 \sim N(900, 0.01)$, $\beta_j \sim N(0, 1)$ and $\tau \sim U(0, 1)$. We also assumed *prior* independence among the parameters. Using the `R` software ([R Core Team, 2015]) and the `R2jags` package, we simulated a total of $121,000$ Gibbs samples (the first $11,000$ samples considered as a burn-in-sample deleted to eliminate the effect of the initial values and using $1,000$ final samples chosen from every 100^{th} sample) to get the *posterior* summaries of interest. Convergence of the simulation algorithm was verified from trace plots of the generated samples for each parameter. Table 15 shows the *posterior* summaries of interest. From the results, we observe that the covariates years and months (linear effects and quadratic effects) show significant effects on the mean of the Normal distribution (positive linear effects of years, positive linear effects on months and negative quadratic effects of months) since the value zero is not inside the 95% credible interval for the regression parameters β_j, $j = 1, 2, 3$.

Table 15 – *Posterior* summaries (daily mean air pressure data)

Parameter	Estimate	Std. Dev.	95% Cred. Int.	
			Lower	Upper
ζ_0	920.7825	0.1509	920.4821	921.0780
ζ_1	0.0604	0.0085	0.0434	0.0772
ζ_2	2.4654	0.0472	2.3736	2.5584
ζ_3	-0.1893	0.0035	-0.1963	-0.1825
τ	0.1304	0.0025	0.1257	0.1353

Daily mean temperature

Considering now, the daily mean temperature data in São Paulo city in the same period of time $(2007 - 2021)$ we also assumed a Normal distribution $N(\mu, \sigma^2)$ in the data analysis, motivated by the histogram of the daily mean air pressure presented in Figure 15. Under a Bayesian approach we assumed the following *prior* (data not considering the presence of covariates) distributions for the parameters of the model: $\mu \sim N(22, 0.001)$ and $\tau = 1/\sigma^2 \sim U(0, 1)$. From the `R` software ([R Core Team, 2015]), (burn-in sample=$1,000$; $1,000$ additional Gibbs samples (every 10^{th} from $10,000$ generated samples) we obtained the Monte Carlo estimates for the *posterior* means μ and τ (between parentheses the

corresponding 95% credible intervals) given, respectively by 20.37 (20.28; 20.45) and 0.09009 (0.08683; 0.09316). That is, the variance σ^2 is estimated by 11, 10 and the standard-deviation σ is estimated by 3.3317. The sample average and sample-deviation of the daily mean temperature data are given, respectively, by, 20.37 and 3.3317.

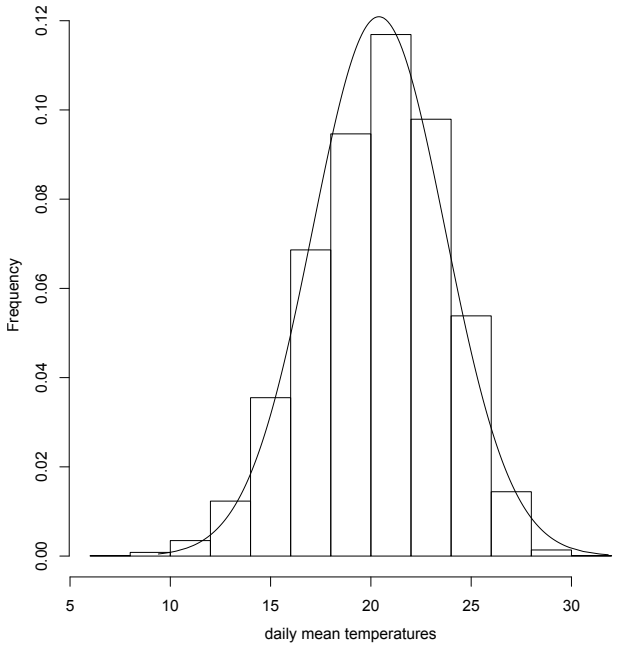

Figure 15 – Histogram of the data (daily mean temperature) and the fitted Normal density with parameters $\mu = 20.37$ and $\sigma = 3.3317$

Figure 15 shows the histogram of the daily mean mean temperature and the fitted Normal density with parameters $\mu = 20.37$ and $\sigma = 3.3317$. We observe a very good fit of the Normal distribution to the daily mean temperature.

Assuming the presence of covariates years (linear effects) and months (linear and quadratic effects) for the daily mean air pressure data we considered the same regression model 5.12 in the mean μ of the Normal distribution $N(\mu, \sigma^2)$. We assumed the following *prior* distributions for the parameters β_0, β_j, $j = 1, 2, 3$ and $\tau = 1/\sigma^2$: $\beta_0 \sim N(22, 0.01)$, $\beta_j \sim N(0, 1)$ and $\tau \sim U(0, 1)$. We also assumed *prior* independence among the parameters. Using the `R` software ([R Core Team, 2015]), we simulated a total of 121, 000 Gibbs samples (the first 11, 000 samples considered as a burn-in-sample deleted to eliminate the effect of the initial values and using 1, 000 final samples chosen from every 100^{th} sample) to get the *posterior* summaries of interest. Convergence of the simulation algorithm was verified from trace plots of the generated samples for each parameter. Table 16 shows the *posterior* summaries of interest. From the results, we observe that the covariates years and months (linear effects and quadratic effects) show significant effects on the mean of the Normal

distribution (positive linear effects of years, negative linear effects on months and positive quadratic effects of months) since the value zero is not inside the 95% credible interval for the regression parameters β_j, $j = 1, 2, 3$.

Table 16 – *Posterior* summaries (the daily mean temperatures data)

Parameter	Estimate	Std. Dev.	95% Cred. Int.	
			Lower	Upper
ζ_0	26.3464	0.1350	26.1192	26.6146
ζ_1	0.0412	0.0082	0.0295	0.0568
ζ_2	-2.4102	0.0499	-2.4927	-2.3113
ζ_3	0.1729	0.0039	0.1662	0.1799
τ	0.1348	0.002	0.1303	0.1395

Daily mean humidity

Considering now the daily mean humidity data in São Paulo city in the same period of time $(2007 - 2021)$ we assumed a Weibull distribution with density,

$$f(t) = \frac{\alpha}{\beta^\alpha} t^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\}, t \geq 0 \tag{5.13}$$

In the data analysis, motivated by the histogram of the daily mean humidity presented in Figure 11. Under a Bayesian approach we assumed the following *prior* (data not considering the presence of covariates) distributions for the parameters of the model:$\alpha \sim U(0.1, 100)$ and $\beta \sim U(0.1, 1000)$. From the `R` software ([R Core Team, 2015]), (burn-in sample=$11,000$; $1,000$ additional Gibbs samples (every 100^{th} from $100,000$ generated samples) we obtained the Monte Carlo estimates for the *posterior* means of α and β (between parentheses the corresponding 95% credible intervals) given, respectively by 7.822 $(7.672; 7.976)$ and 74.59 $(74.33; 74.87)$.

Figure 16 shows the histogram of the daily mean humidity and the fitted Weibull density with parameters $\alpha = 7.822$ and $\beta = 74.59$. From this figure we observe a very good fit of the Weibull distribution to the dataset (daily mean humidity).

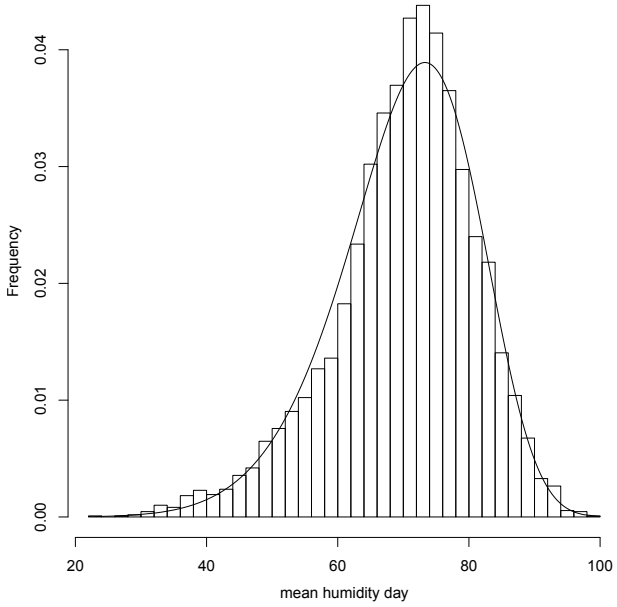

Figure 16 – Histogram of the data (daily mean humidity) and the fitted Weibull density with parameters $\alpha = 7.822$ and $\beta = 74.59$

Assuming the presence of covariates years (linear effects) and months (linear and quadratic effects) for the daily mean humidity data we considered the following regression model in the mean β of the Weibul distribution $Wei(\alpha, \beta)$:

$$\beta_i = \gamma_0 + \gamma_1(years_i) + \gamma_2(months_i) + \gamma_3(months_i)^2 \tag{5.14}$$

where $i = 1, 2, \ldots, 5479$ (number of days from January 01, 2007 to 31 December, 2021).

Table 17 – *Posterior* summaries (the daily mean humidity data)

Parameter	Estimate	Std. Dev.	95% Cred. Int.	
			Lower	Upper
α	7.9453	0.0838	7.7823	8.1094
γ_0	4.3793	0.0070	4.3656	4.3930
γ_1	-0.0055	0.0004	-0.0062	-0.0047
γ_2	-0.0078	0.0022	-0.0121	-0.0035
γ_3	0.0005	0.0002	0.0002	0.0008

We assumed the following *prior* distributions for the parameters γ_0, γ_j, $j = 1, 2, 3$ and α: $\gamma_0 \sim N(3, 0.1)$, $\gamma_j \sim N(0, 1)$ and $\alpha \sim Gamma(60, 10)$, where $Gamma(a, b)$, denotes a Gamma distribution with mean a/b and variance a/b^2. We also assumed *prior* independence among the parameters. Using the `R` software [R Core Team, 2015], we

simulated a total of 411, 000 Gibbs samples (the first 211, 000 samples considered as a burn-in-sample deleted to eliminate the effect of the initial values and using 20, 000 final samples chosen from every 10^th sample) to get the *posterior* summaries of interest. Convergence of the simulation algorithm was verified from trace plots of the generated samples for each parameter. Table 17 shows the *posterior* summaries of interest. From the results, we observe that the covariates years and months (linear effects and quadratic effects) show significant effects on the parameter β of the Weibull distribution (negative linear effects of years, negative linear effects on months and positive quadratic effects of months) since the value zero is not inside the 95% credible interval for the regression parameters γ_j , $j = 1, 2, 3$ (or for the means $E(T_i) = \beta_i\Gamma(1 + 1/\alpha)$ of the Weibull distribution given the values of the covariates).

5.5 Concluding remarks

The results of this study showed that the use of a Tobit-Weibull model under a Bayesian approach can be useful for analyzing daily rainfall data, as observed for data from the city of São Paulo, Brazil. Other climate variables such as mean daily temperature, mean daily air pressure and mean daily humidity for the city of São Paulo from 2007 to 2021 were also analyzed under a Bayesian approach using traditional models as regression model with normal errors (temperature and air pressure) and Weibull regression for asymmetric data (air humidity).

The inferences of interest were obtained from the simulation of samples for the joint *posterior* distribution of the parameters of each model using MCMC simulation methods that were greatly facilitated using the `R` software [R Core Team, 2015] which does not require much computation effort to obtain the a *posterior* summaries of interest.

Some important conclusions were obtained from the data analysis:

- Considering the daily precipitation data, we observed that the covariate months (linear effects) show significant effects on the scale of the Weibull distribution (negative linear effects on months, that is, decreasing in daily total precipitations in the last months of the year) since the value zero is not inside the 95% credible interval for the regression parameter γ_2; the same conclusions are observed for the probabilities to have days with rain using a logistic regression model.
- Considering the daily mean temperatures, we observed that the covariates years

(linear effects) and months (linear effects and quadratic effects) show significant effects on the mean of the Normal distribution (positive linear effects of years, negative linear effects on months and positive quadratic effects of months) since the value zero is not inside the 95% credible interval for the regression parameters β_j, $j = 1, 2, 3$.

- Considering the daily mean air pressures, we observed that the covariates years (linear effects) and months (linear effects and quadratic effects) show significant effects on the mean of the Normal distribution (positive linear effects of years, negative linear effects on months and positive quadratic effects of months) since the value zero is not inside the 95% credible interval for the regression parameters β_j, $j = 1, 2, 3$.

- Considering the daily mean air humidity, we observed that the covariates years and months (linear effects and quadratic effects) show significant effects on the parameter β of the Weibull distribution (negative linear effects of years, negative linear effects on months and positive quadratic effects of months) since the value zero is not inside the 95% credible interval for the regression parameters γ_j, $j = 1, 2, 3$ (or for the means $E(T_i) = \beta_i\Gamma(1 + 1/\alpha)$ of the Weibull distribution given the values of the covariates).

Chapter 6

Tobit-generalized Weibull models under a Bayesian approach applied to daily rain precipitation data

6.1 Introduction

In this chapter we will use the same dataset as in the previous chapter with a different approach. The daily rainfall data have excess zeros, i.e. many days without rain, which requires special statistical models for data analysis. Two generalized forms of the Weibull distribution will be used: the exponentiated Weibull (EW) model [Mudholkar and Srivastava, 1993] and the generalized modified Weibull model [Carrasco et al., 2008] assuming a Tobit structure. Given the large number of zero values (days without rainfall), the proposed models will be used as alternatives to the standard Weibull model (with two parameters) in the presence of left-censored data.

The paper is organized as follows: section 2 presents the Tobit model assuming a Weibull distribution and generalizations of the Weibull distribution; section 3 presents the data set; section 4 presents the Bayesian results obtained; finally section 5 presents some concluding remarks.

6.2 Materials and Methods

Mixture model

If we have a complete observation, that is, $(T > C)$, in this work we assume a truncated generalized form of the Weibull distribution for the random variable T with probability density function given by,

$$f(t \mid T > C) = \frac{f_0(t)}{P(T > C)} \tag{6.1}$$

where

$$S_0(t) = P(T > t) = \exp\left\{-\left(\frac{t}{\beta}\right)^{\alpha}\right\}$$

Let us assume a mixture model, given by the probability density function,

$$f(t) = p[dC(t)] + (1-p)\frac{f_0(t)}{S_0(C)} \tag{6.2}$$

where $dC(t)$ is the Dirac measure at $(0, C)$ (in the mixture model, we are assuming a degenerated probability distribution $P(T = 0) = 1$).

The likelihood function for the parameters p, α and β based on one observation t is given from (5.2), by

$$L(p, \alpha, \beta) = p[dC(t)] + (1-p)\frac{f_0(t)}{S_0(C)}$$

where $f_0(t) = \frac{\alpha}{\beta^{\alpha}} t^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^{\alpha}\right\}$ and $S_0(C) = \exp\left\{-\left(\frac{C}{\beta}\right)^{\alpha}\right\}$.

With the censoring information (see (5.1)), let us define a binary variable $\delta = 1$ if T is a complete observation $(T > C)$ and $\delta = 0$ if T is a left censored observation $(T \leq C)$ with conditional probabilities given by

$$\begin{aligned} P(\delta = 0 \mid p, \alpha, \beta, t) &= \frac{p}{p+(1-p)\frac{f_0(t)}{S_0(C)}} \\ \text{and,} & \\ P(\delta = 1 \mid p, \alpha, \beta, t) &= \frac{(1-p)\frac{f_0(t)}{S_0(C)}}{p+(1-p)\frac{f_0(t)}{S_0(C)}} \end{aligned}$$

In this way, we have a Bernoulli distribution. The likelihood function $L(p, \alpha, \beta)$ based on n observations is given by

$$L(p, \alpha, \beta; \mathbf{t}) = \prod_{i=1}^{n} p^{(1-\delta_i)} \left[(1-p)\frac{f_0(t_i)}{S_0(C)}\right]^{\delta_i} \tag{6.3}$$

Remark: For the daily precipitation data we have $\delta = 1$ (day with rain precipitation) or $\delta = 0$ (day without rain precipitation, that is, $T = 0$).

The Tobit-generalized modified Weibull model (TGMW)

A generalized modified Weibull (GMW) distribution with four parameters for the amount of daily rain precipitation is defined by a probability density function given by,

$$f(t) = \frac{\alpha\beta t^{\gamma-1}(\gamma + \lambda t)\exp\left(\lambda t - \alpha t^{\gamma}e^{\lambda t}\right)}{\{1 - \exp\left(-\alpha t^{\gamma}e^{\lambda t}\right)\}^{(1-\beta)}} \tag{6.4}$$

where $t > 0$, α, β, γ and λ are positive parameters and the survival function $S(t) = P(T > t)$ is given by,

$$S(t) = 1 - \left\{1 - \exp\left(-\alpha t^{\gamma}e^{\lambda t}\right)\right\}^{\beta} \tag{6.5}$$

From the equations 6.3, 6.4 and 6.5, the likelihood function, for n observations, for the Tobit-GMW model is given by

$$L(p, \alpha, \beta, \gamma, \lambda; \mathbf{t}) = \prod_{i=1}^{n} p^{(1-\delta_i)}\left[(1-p)\frac{\frac{\alpha\beta t_i^{\gamma-1}(\gamma+\lambda t_i)\exp\left(\lambda t_i - \alpha t_i^{\gamma}e^{\lambda t_i}\right)}{\left\{1-\exp\left(-\alpha t_i^{\gamma}e^{\lambda t_i}\right)\right\}^{(1-\beta)}}}{1 - \{1 - \exp\left(-\alpha C^{\gamma}e^{\lambda C}\right)\}^{\beta}}\right]^{\delta_i} \tag{6.6}$$

The Tobit-exponentiated Weibull model (TEW)

A exponentiated Weibull (EW) distribution with three parameters for the amount of daily rain precipitation is obtained from (6.4) assuming $\lambda = 0$, with probability density function given by,

$$f(t) = \frac{\alpha\beta\gamma t^{\gamma-1}\exp\left(-\alpha t^{\gamma}\right)}{\{1 - \exp\left(-\alpha t^{\gamma}\right)\}^{1-\beta}} \tag{6.7}$$

where $t > 0$, α, β and γ are positive parameters and the survival function $S(t) = P(T > t)$ is given by,

$$S(t) = 1 - \{1 - \exp\left(-\alpha t^{\gamma}\right)\}^{\beta} \tag{6.8}$$

From the equations 6.3, 6.7 and 6.8, the likelihood function, for n observations, for the Tobit-EW model is given by

$$L(p,\alpha,\beta,\gamma;\mathbf{t}) = \prod_{i=1}^{n} p^{(1-\delta_i)} \left[(1-p) \frac{\frac{\alpha\beta\gamma t_i^{\gamma-1}\exp\left(-\alpha t_i^{\gamma}\right)}{\left\{1-\exp\left(-\alpha t_i^{\gamma}\right)\right\}^{1-\beta}}}{1-\left\{1-\exp\left(-\alpha C^{\gamma}\right)\right\}^{\beta}} \right]^{\delta_i} \tag{6.9}$$

The Tobit-Weibull model (TW)

A popular probability distribution for the amount of daily rain precipitation is obtained from (6.4) assuming $\lambda = 0$ and $\beta = 1$ (a Weibull distribution with two parameters), with probability density function given by,

$$f(t) = \alpha\gamma t^{\gamma-1}\exp\left(-\alpha t^{\gamma}\right) \tag{6.10}$$

where $t > 0$, α and γ are positive parameters and the survival function $S(t) = P(T > t)$ is given by,

$$S(t) = \exp\left(-\alpha t^{\gamma}\right) \tag{6.11}$$

From the equations 6.3, 6.10 and 6.11, the likelihood function, for n observations, for the Tobit-Weibull model is given by

$$L(p,\alpha,\beta,\gamma;\mathbf{t}) = \prod_{i=1}^{n} p^{(1-\delta_i)} \left[(1-p) \frac{\alpha\gamma t_i^{\gamma-1}\exp\left(-\alpha t_i^{\gamma}\right)}{\exp\left(-\alpha C^{\gamma}\right)} \right]^{\delta_i} \tag{6.12}$$

Inferences for the parameters of the models are obtained under a Bayesian approach using existing MCMC methods like the Metropolis-within-Gibbs algorithms. In this way, in the simulation of samples of the joint *posterior* distribution [Gelfand and Smith, 1990], $\pi(\theta \mid data)$ where θ is the vector of all parameters, we use Gibbs or Metropolis-Hastings algorithms, where it is needed to sample each parameter from the *posterior* conditional distributions $\pi(\theta_r \mid \theta_{(r)}, data)$, where $\theta_{(r)}$ denotes the vector of all parameters except θ_r and r is associated to each one of the parameters of the model. To simplify the computational work in the iterative procedure to get the Bayesian inferences, the literature presents different free softwares to simulate samples of the joint *posterior* distribution of interest.

For a Bayesian analysis we assume uniform $U(a,b)$ *prior* probability distributions for the parameters of the proposed generalized forms of the Weibull distribution assuming a and b known hyperparameters and a $Beta(e,f)$ or a $U(0,1)$ *prior* probability distribution for p with e and f known hyperparameters in the Tobit form of the model. We further assume *prior* independence among the three parameters.

In presence of a vector of covariates $\mathbf{x} = (x_1, x_2, ..., x_p)^\top$ let us assume a logit model for the parameter p, given by,

$$logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \zeta_0 + \zeta_1 x_1 + \zeta_2 x_2 + \ldots + \zeta_p x_p \tag{6.13}$$

For a Bayesian analysis it assumed a uniform $U(a, b)$ *prior* distribution for the parameter α and Normal $N(c, d^2)$ *prior* distributions for the regression parameters $\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_p$ in (6.13) with c and d known hyperparameters and Normal $N(e, f^2)$ *prior* distributions for the regression parameters $\zeta_0, \zeta_1, \zeta_2, \ldots, \zeta_p$ with e and f known hyperparameters. We use MCMC simulation methods to get *posterior* summaries of interest for the parameters of the models (see, for example, Chib and Greenberg [1995], Gelfand and Smith [1990], Gelman et al. [1995], Geman and Geman [1984], Gilks et al. [1995]).

6.3 Application to a climatic dataset

The original climatic data considered in this study (total rainfall per hour in mm; atmospheric pressure at the station level in mB; temperature per hour in ^oC; relative air humidity per hour in %) was obtained from the Mirante climate station located in the city of São Paulo, Brazil (latitude = -23.59; longitude = -46.52 and altitude = 785.64 meters) for the period from January 1, 2007 to December 31, 2021 (n = 5479 measurements). Dataset obtained in https://portal.inmet.gov.br/dadoshistoricos. This site presents climate data for several cities in Brazil. Missing observations were imputed as means (previous and *posterior* observed values) for the variables temperature and humidity; for total missing precipitation, the value zero was considered. The possible limitation of the data is related to an apparently short period (15 years) as many weather stations in Brazil given by INMET (Instituto Nacional de Meteorologia, Brasil) were created recently, a common fact in third world countries where it is usually difficult to obtain longer series of climatic observations.

The dataset contains 2323 days with rain and 3156 days without rain (total of n = 5479 days). It is important to point out that we could assume in a first modeling approach, the zero data (days with no rain) as left censored data with the same probability distribution as the complete data, but this approach could lead to very poor inferences when there is a great proportion of left censored data as in our case of daily precipitation data.

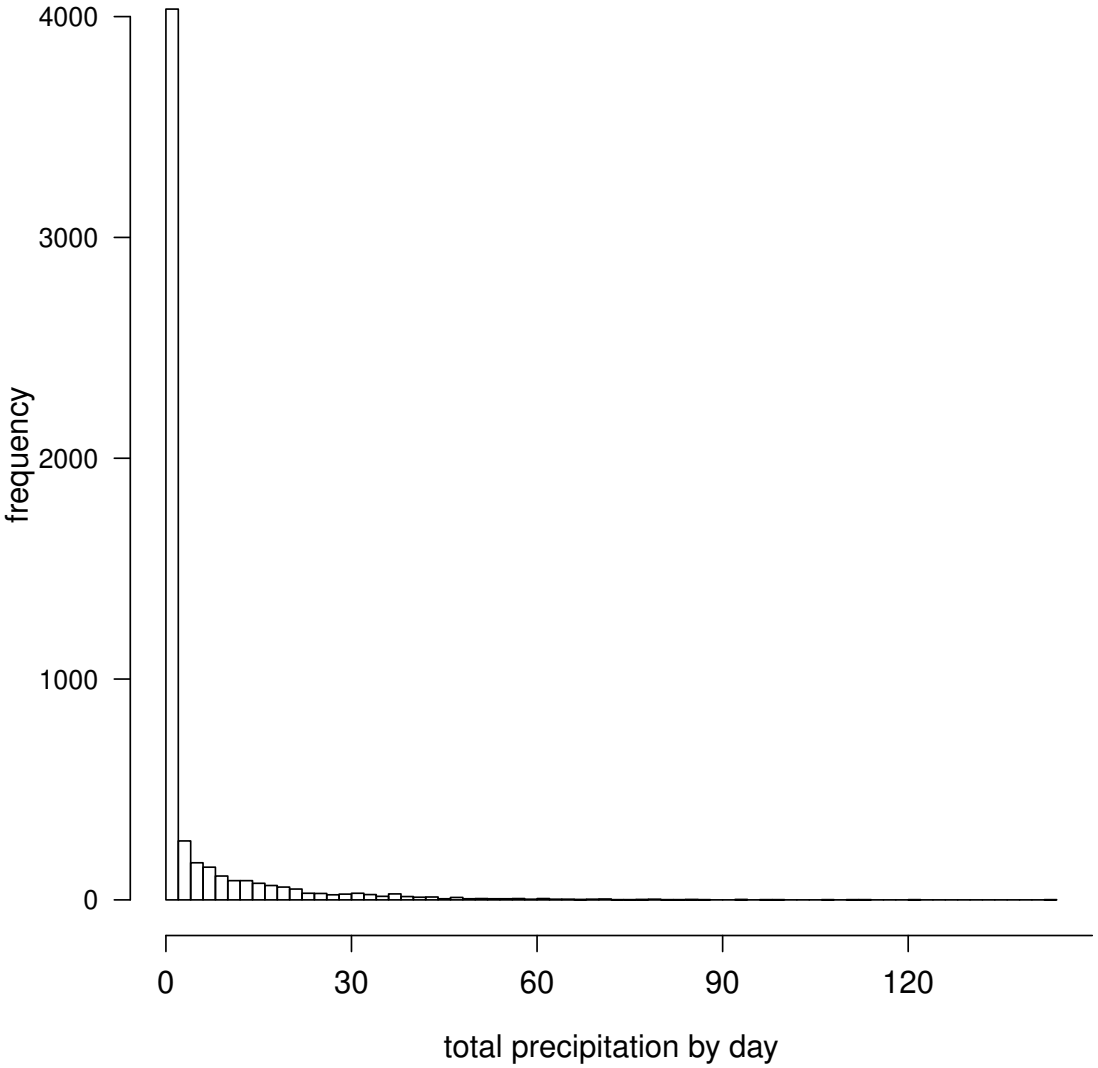

Figure 17 – Histogram of daily amount of rain (2323 days) in São Paulo city for the period 2007-2021

Figure 17 shows the histogram of daily amount of rain (2323 days) in São Paulo city for the period 2007-2021. We observe an asymmetric form for the histogram indicating the need for a probability distribution for continuous positive data that captures the asymmetric behavior of the dataset.

We define an indicator of censoring variable, $\delta = 1$ if a positive random variable T has a complete observation $(T > C)$ and $\delta = 0$ if T is a left censored observation $(T \leq C)$. Here we could assume as left-censoring considering, as a special case, the daily rain precipitation data with the zero value (no precipitation in a day) and C is arbitrary value fixed as the value 0.01, as a special case.

6.4 Results

In this section we present the results of a Bayesian analysis for the daily precipitation in São Paulo from 2007 to 2021, considering the Tobit model assuming the three probability distributions introduced in section 6.2. The statistical analysis was carried out in the `R` software [R Core Team, 2015] and the `R2jags` package was used to obtain the Bayesian estimates for the models parameters.

Assuming Tobit-Weibull, Tobit-Exponentiated Weibull (TEW) and Tobit-generalized modified Weibull (TWG) mixture models for the daily precipitation data not considering the presence of covariates, we simulated Gibbs samples, for each model, using final sample sizes of $1,000$ to obtain the summaries of interest from the joint parameter distribution, which are available in Table 18. Note that in all three models, the estimates for the mixture part, $\hat{p} = 0.5760$, exactly the proportion of days without rain ($3156/5479 = 0.5760$).

Table 18 – *Posterior* summaries for the Tobit-Weibull, Tobit-Exponentiated Weibull and Tobit-modified generalized Weibull models for the daily precipitation data not considering the presence of covariates

Model	Parameter	Estimate	Std. Dev.	95% Cred. Int.	
				Lower	Upper
TW	α	0.2680	0.0096	0.2516	0.2884
	γ	0.6537	0.0111	0.6309	0.6762
	p	0.5766	0.0064	0.5635	0.5875
TEW	α	0.5135	0.1328	0.2980	0.8081
	β	1.5737	0.3616	1.0399	2.4326
	γ	0.2082	0.0168	0.4086	0.6246
	p	0.5760	0.0065	0.5634	0.5887
TGMW	α	1.9067	0.1883	1.5180	2.2625
	β	8.9640	1.8324	5.7285	12.9800
	γ	0.2082	0.0168	0.1789	0.2457
	λ	0.0053	0.0005	0.0044	0.0062
	p	0.5760	0.0068	0.5633	0.5894

In constructing the *posterior* summaries, we considered the following *prior* distributions: $\alpha \sim U(0,1), \gamma \sim U(0,2)$ and $p \sim U(0,1)$ for the parameters for Tobit-Weibull model; $\alpha \sim U0,1), \beta \sim U(0,100), \gamma \sim U(0,2)$ and $p \sim U(0,1)$ for the parameters for TEW model and $\alpha \sim G(1,2), \beta \sim G(3,2), \gamma \sim G(1,2), \lambda \sim U(0,0.1)$ and $p \sim U(0,1)$ for the parameters for TGW model. We also assumed *prior* independence among the parameters.

Use of Tobit structure

The Tobit structure for the daily precipitation data in presence of covariates years (linear effects) and months (linear and quadratic effects) is gives by

$$logit(p) = \zeta_0 + \zeta_1(years_i) + \zeta_2(months_i) + \zeta_3(months_i)^2 \tag{6.14}$$

where $i = 1, 2, \ldots, 5479$ (number of days from January 01, 2007 to 31 December, 2021).

Assuming Tobit structure in the Weibull, Exponentiated Weibull and modified generalized Weibull models for the daily precipitation data, we first simulated, for each model, a total of 110, 000 Gibbs samples considered as a burn-in-sample deleted to eliminate the effect of the initial values and using 1, 000 final samples to get the *posterior* summaries of interest. We assume the following *prior* distributions for the models parameters:

- $\alpha \sim U(0, 10)$, $\gamma \sim U(0, 2)$, $\zeta_0 \sim N(0, 1)$ and $\zeta_j \sim N(0, 100)$, $j = 1, 2, 3$ for Tobit Weibull model;
- $\alpha \sim U(0, 100), \beta \sim U(0, 20000), \gamma \sim U(0, 2), \zeta_0 \sim N(0, 1)$ and $\zeta_j \sim N(0, 100)$, $j = 1, 2, 3$, for Tobit-Exponentiated Weibull model and
- $\alpha \sim U(0, 100), \beta \sim U(0, 100000), \gamma \sim U(0, 2), \lambda \sim U(0, 10), \zeta_0 \sim N(0, 1)$ and $\zeta_j \sim N(0, 100)$, $j = 1, 2, 3$ Tobit-modified generalized Weibull.
- We also assumed *prior* independence among the parameters.

Table 19 – *Posterior* summaries for the Tobit-Weibull, Tobit-Exponentiated Weibull and Tobit-modified generalized Weibull models for the daily precipitation data.

Model	Parameter	Estimate	Std. Dev.	95% Cred. Int.	
				Lower	Upper
TW	α	0.2704	0.0099	0.2502	0.2896
	γ	0.6498	0.0113	0.6279	0.6735
	ζ_0	-0.2926	1.3578	-3.6535	1.7215
	ζ_1	-0.0006	0.0007	-0.0016	0.0010
	ζ_2	0.7150	0.0329	0.6484	0.7783
	ζ_3	-0.0514	0.0024	-0.0562	-0.0467
TEW	α	0.4996	0.1240	0.2894	0.7652
	β	1.5332	0.3317	1.0076	2.2888
	γ	0.5189	0.0548	0.4227	0.6340
	ζ_0	-0.2834	1.0037	-2.2758	1.6218
	ζ_1	-0.0006	0.0005	-0.0016	0.0004
	ζ_2	0.7120	0.0340	0.6431	0.7786
	ζ_3	-0.0512	0.0025	-0.0560	-0.0461
TGMW	α	9.0326	0.0407	8.9288	9.0966
	β	9621.2129	333.5861	8746.4723	9988.5013
	γ	0.0598	0.0010	0.0579	0.0619
	λ	0.0028	0.0002	0.0024	0.0032
	ζ_0	-3.5043	2.2515	-6.9452	-0.4028
	ζ_1	0.0015	0.0011	0.0001	0.0032
	ζ_2	0.4139	0.0316	0.3560	0.4801
	ζ_3	-0.0317	0.0024	-0.0367	-0.0271

Table 19 also shows the *posterior* summaries of interest. From the results, we observe that the covariate months (linear effects) show a significant effect on the probabilities of having days with rain, since the zero value is not within the 95% credible interval for the regression parameter ζ_2; the same conclusions are observed for the quadratic effects of the mont

Discrimination of the proposed models

A Bayesian discrimination method can be used to compare the three forms of the Tobit models (Tobit-Weibull, Tobit EW, and Tobit GMW models). Thus, to select the best model, we consider the use of the *posterior* Bayes factor and use the Gibbs samples generated for the parameters of each model to obtain Monte Carlo estimates of the Bayes

factor.

The *posterior* Bayes factor is as a discrimination criterion between two models i and j given by $B_{ij} = V_i/V_j$ where V_k is the *posterior* mean of the likelihood function under model k given by

$$V_k = \int L(D \mid \theta_k) P(\theta_k \mid D) d\theta_k \tag{6.15}$$

where $L(D \mid \theta_k)$ is the likelihood function under model k and $P(\theta_k \mid D)$ is the joint *posterior* distribution of the vector of parameters θ_k. If $B_{ij} = V_i/V_j > 1$, then the Bayes factor criterion favors model i.

We use the Monte Carlo estimation of the expected value of the likelihood function (or the log-likelihood function) for each model. That would correspond to the values V_i given in (6.15). Once the values of V_i are obtained for each model, $i = 1, 2$, the quantity $B_{ij} = V_i/V_j$, may also be obtained. Assuming the Tobit-Weibull (model 1) and Tobit-EW (model 2) not considering the presence of covariates, we obtain $B_{12} = 0.9952$. Assuming the Tobit-Weibull (model 1) and Tobit-GMW (model 3), we get $B_{13} = 0.9852$ and assuming Assuming the Tobit-EW (model 2) and Tobit-GMW (model 3), we get $B_{23} = 0.99$, an indication that the three models have similars fit for the data in presence of left-censored data. Therefore, we can conclude that the use of Tobit Weibull model (model 1) is preferable to the other two models considered (Tobit-EW and Tobit-GMW).

Comparing the three proposed models in presence of the covariates, we get $B_{12} = 0.9979$, an indication that both models 1 and 2, (Tobit-Weibull model) and (Weibull-EW model) have similar fit for the data in presence of left-censored data. In the same way, we get $B_{13} = 1.9007$, an indication that model 1 (Tobit-Weibull model) is better fitted by the data when compared to model 3 (Weibull-GMW model) in presence of left-censored data) and $B_{23} = 1.9048$, an indication that model 2 (Weibull-EW model) is better fitted by the data when compared to model 3 (Tobit-GMW model) in presence of left-censored data.

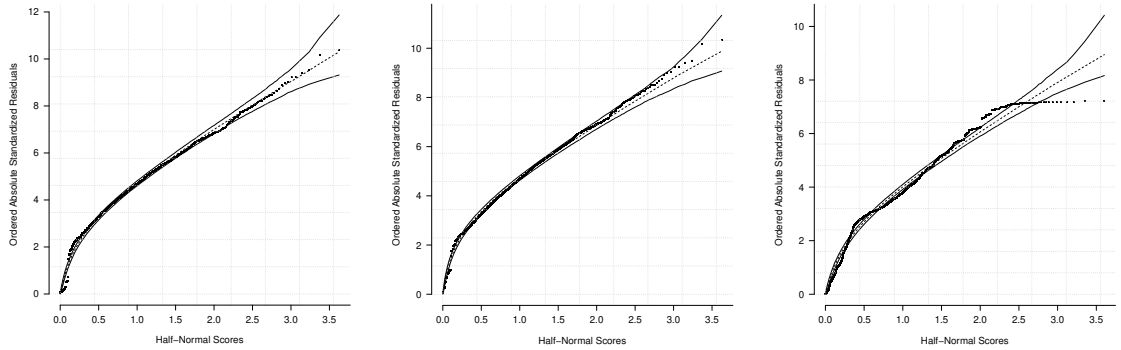

Figure 18 – Envelope for the residuals in presence of covariate and left-censored data for the model 1 – Tobit-Weibull, model 2 – Tobit-EW and model 3 – Tobit-GMW.

Figure 18 shows the Half-Normal plot with simulated envelope of the residual for the three models, in presence of covariates and of left-censored data. The predictions obtained from the model 3 could be inaccurate since there are many non-significant factors that could disturb the model estimates. This may be related to the fact that there is the presence of about 57.6% left censored observations. However, the model 1 the majority of the observed values are within the 95% credible range for the proposed regression model which is an indication of good accuracy.

6.5 Concluding remarks

In this study we explored the use of Tobit models assuming generalized forms of the Weibull distribution, a model widely used in the area of reliability in engineering and industry or in the area of survival analysis to analyze daily rainfall data. Daily rainfall data in general have many days without the presence of rain, that is, zero observations, indicating the need for specific models that incorporate the excess of zeros. In particular, we used and compared three special models in the analysis of rainfall data in the city of São Paulo, Brazil under a Bayesian approach. The use of the three models without considering the presence of covariates led to similar Bayesian results from which it is concluded that the use of a Weibull model that only has two parameters (parsimony) is preferable to the other two models considered (Tobit-EW and Tobit-GMW).

It is important to point out that some recent generalizations of the Weibull distribution with three or four parameters introduced in the literature show non-identifiability problems leading to instability in the determination of usual point estimators for the parameters of the models under classical or Bayesian approaches. These problems were

observed in the Bayesian analysis of the daily rain precipitation data of São Paulo city, where the convergence of the MCMC simulation algorithm used to simulate samples of the joint *posterior* distribution for the parameters of the Tobit-GMW model only was obtained using very informative *prior* distributions for the parameters of the model from information of the obtained results assuming the Tobit-EW model. For the other two proposed models (Tobit-Weibull and Tobit-EW models) the convergence of the simulation algorithm was easily obtained assuming approximately non-informative *prior* distributions. From these results, we must emphasize that the use of generalized models such as the assumed Tobit-GMW model in the analysis of daily precipitation data obtained from the the city of São Paulo must be done carefully. Assuming the three proposed models, we observed that the Bayesian estimators for p (probability of day without rain) are very similar and close to the empirical estimator given by the ratio of the number of days with no rain over the total observed number of days ($3176/5479 = 0,5760$).

In presence of covariates we found similar inference results assuming the Tobit-Weibull and Tobit-EW models with the covariate months (linear effect) in the logistic regression model showing significant effect on the amount of daily rain precipitation (95% credible interval does not contain the zero value); assuming the Tobit-GMW model we also found that the covariate month (linear and quadratic effects) shows significant effect on on the amount of daily rain precipitation. As concluding remarks, it is important to point out that the use of the proposed Tobit model assuming the Weibull or generalized forms of the Weibull distribution for the statistical analysis of the amount of daily rain precipitation could be very useful in climate data studies, with simple interpretations and simple computational work to get the *posterior* sumaries of interest under a Bayesian approach especially using the free existing `R` software (R Core Team, 2015). A weak point of the obtained results in the application of the proposed methodology to daily precipitation in São Paulo city could be the short time of follow-up (January 2007 to December, 2021), showing that the covariate year do not show significant effect on the probability of daily rain precipitation in São Paulo city. Better results could be obtained assuming longer follow-up periods.

Chapter 7

General Conclusions

The search of appropriate probability distributions for data analysis still is a great problem in most studies, especially assuming the left-censoring data structure. In this thesis, it was presented some techniques to model this kind of data based on Tobit and Weibull structure. Initially, we worked with environmental data related to ammonia nitrogen concentrations in U.S. rivers where we introduced a univariate Tobit-Weibull based on a mixture approach. Some properties of this new distribution were also discussed in this study and an extensive simulation study was performed to verify the effectiveness of the maximum likelihood estimation method assuming different fixed values for the parameters of the model and different sample sizes. The results obtained from Monte Carlo studies showed that the biases and RMSEs of the Tobit-Weibull model are asymptotically non-biased. Also, based on regression structure adopted, we identified important factors according to the literature that affects the ammonia nitrogen concentrations even using non-informative *prior* distributions for the regression parameters which implies the proposed model could be a great alternative for left-censored data analysis.

In a second approach, it was proposed a bivariate Tobit-Weibull model under left-censoring scheme in order to analyzes a stellar data which is common the presence of left-censored data. Since the bivariate model inherits most of properties of the univariate model, we expected that the bivariate as able to identify some covariates assuming non-informative *prior* distributions for the regression structure. Thus, based on the data analysis, we found that the proposed model was accurate to accomplish our goal and do a good prediction, especially for marginal means.

In a third approach, we introduced both proposed models, univariate and bivariate Tobit-Weibull, for the analysis of left-censored medical data related to cancer and vaccines. The obtained results of this study showed many advantages for the use of Tobit-Weibull

models in terms of great accuracy for the obtained point and interval inferences, great computational simplicity to get the inferences of interest under hierarchal Bayesian approach as well simple interpretations for the model parameters.

A fourth approach considered an analysis of rainfall data, where the response variable was the total daily precipitation of a climate station located in the city of São Paulo, Brazil, over the 24-year period (2007 until 2021). We fitted a Tobit-Weibull model in the presence of some covariates (linear effects of years, linear and quadratic effects of months). The results showed that the use of a Tobit-Weibull model under a Bayesian approach can be useful for the analysis of daily rainfall data. We also simultaneously used a logistic regression model for the occurrence (or not) of daily rainfall over the follow-up time period. Other climate variables such as mean daily temperature, mean daily atmospheric pressure and mean daily humidity for the city of São Paulo in the same period were also analyzed under a Bayesian approach using traditional models such as regression with normal errors for data with asymmetry - temperature and air pressure) and Weibull regression for asymmetric data such as air humidity data.

Finally, an approach for data with excess zeros was considered. We explored the use of Tobit models assuming generalized forms of the Weibull distribution, to analyze daily rainfall data. Daily rainfall data in general have many days without the presence of rain, i.e. zero observations, indicating the need for specific models that incorporate excess zeros. Using the three models without considering the presence of covariates led to similar Bayesian results, from which we conclude that using a Weibull model that has only two parameters (parsimony) is preferable to the other two models considered (Tobit-EW and Tobit-GMW). In presence of covariates, we find similar inference results assuming the Tobit-Weibull and Tobit-EW models with the covariate months (linear effect) in the logistic regression model showing a significant effect on the amount of daily rainfall. the Tobit-GMW model, we also find that the covariate month (linear and quadratic effect) shows a significant effect on the amount of daily rainfall. As concluding remarks, it is important to note that the use of the proposed Tobit model assuming the Weibull distribution or generalized forms of the Weibull distribution for the statistical analysis of the amount of daily rainfall could be very useful in climate data studies, with simple interpretations and simple computational work to obtain the subsequent summaries of interest under a Bayesian approach.

In conclusion, the results emerging from this study reinforce the fact that the search of appropriate statistical model could be extremely difficult depending on the censoring structure of the lifetime data. However, the proposed methodology could be very useful in

the medical data analysis in presence of left-censored scheme. In addition, the identification of important covariates was also easily obtained assuming the proposed models even using non-informative *priors* for the parameters of the model, under a hierarchal Bayesian approach. The results could be also extended to other cross-over trials in clinical research; reliability analysis in engineering; risk analysis in economics; among many others areas.

Bibliography

J. Achcar, R. Brookmeyer, and W. Hunter. An application of bayesian analysis to medical follow-up data. *Statistics in Medicine*, 4(4):509–520, 1985. ISSN 0277-6715. doi: 10.1002/sim.4780040411. Copyright: Copyright 2016 Elsevier B.V., All rights reserved.

J. Achcar, V. Cancho, and H. Bolfarine. A bayesian analysis for exponentiated-weibull distribution. *Journal of Applied Statistical Science*, 4(8):227–242, 1999.

J. Achcar, E. Coelho-Barros, J. Cuevas, and J. Mazucheli. Use of Lévy distribution to analyze longitudinal data with asymmetric distribution and presence of left censored data. *Communications for Statistical Applications and Methods*, 25(1):43–60, 2018. doi: 10.29220/CSAM.2018.25.1.043.

J. A. Achcar, E. A. Coelho-Barros, and J. Mazucheli. Cure fraction models using mixture and non-mixture models. *Tatra Mountains Mathematical Publications*, 51(1):1–9, 2012.

J. A. Achcar, E. A. Coelho-Barros, and J. Mazucheli. Block and basu bivariate lifetime distribution in the presence of cure fraction. *Journal of Applied Statistics*, 40(9):1864–1874, 2013.

M. Akritas, T. Ruscitti, and G. Patil. Statistical analysis of censored environmental data. *Handbook of Statistics*, 12:221–242, 12 1994. ISSN 0169-7161. doi: 10.1016/S0169-7161(05)80009-4.

L. V. Alexander, X. Zhang, T. C. Peterson, J. Caesar, B. Gleason, A. M. G. Klein Tank, M. Haylock, D. Collins, B. Trewin, F. Rahimzadeh, A. Tagipour, K. Rupa Kumar, J. Revadekar, G. Griffiths, L. Vincent, D. B. Stephenson, J. Burn, E. Aguilar, M. Brunet, M. Taylor, M. New, P. Zhai, M. Rusticucci, and J. L. Vazquez-Aguirre. Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research: Atmospheres*, 111(D5), 2006. doi: https://doi.org/10.1029/2005JD006290. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JD006290.

T. Amemiya. Tobit models: A survey. *Journal of Econometrics*, 24(1):3 – 61, 1984. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(84)90074-5.

R. Arellano-Valle, L. Castro, G. González Farías, and K. Munõz Gajardo. Student-t censored regression model: Properties and inference. *Statistical Methods & Applications*, 21(4):453–473, 2012. doi: 10.1007/s10260-012-0199-y.

B. Arnell, N. andLloyd-Hughes. The global-scale impacts of climate change on water resources and flooding under new climate and socio-economic scenarios. *Climatic Change*, 122(1):127–140, January 2014. doi: 10.1007/s10584-013-0948-4. URL https://ideas.repec.org/a/spr/climat/v122y2014i1p127-140.html.

B. C. Arnold and D. Strauss. Bivariate distributions with exponential conditionals. *Journal of the American Statistical Association*, 83(402):522–527, 1988.

C. Atwood, L. Blackwood, G. Harris, and C. Loehr. Recommended methods for statistical analysis of data containing less-than-detectable measurements. 1991. doi: 10.2172/6174750.

B. H. Auestad, A. Henriksen, and H. A. Karlsen. *Modeling and Analysis of Daily Rainfall Data*, pages 493–503. Springer Netherlands, Dordrecht, 2012.

N. Balakrishnan. Approximate mle of the scale parameter of the rayleigh distribution with censoring. *IEEE Transactions on Reliability*, 38(3):355–357, 1989.

N. Balakrishnan and J. Varadan. Approximate mles for the location and scale parameters of the extreme value distribution with censoring. *IEEE Transactions on Reliability*, 40 (2):146–151, 1991.

A. Bárdossy, E. Modiri, F. Anwar, and G. Pegram. Gridded daily precipitation data for iran: A comparison of different methods. *Journal of Hydrology: Regional Studies*, 38: 100958, 2021. ISSN 2214-5818. doi: https://doi.org/10.1016/j.ejrh.2021.100958. URL https://www.sciencedirect.com/science/article/pii/S2214581821001877.

M. Barros, M. Galea, M. González, and V. Leiva. Influence diagnostics in the tobit censored response model. *Statistical Methods and Applications*, 19:379–397, 08 2010. doi: 10.1007/s10260-010-0135-y.

M. Barros, M. Galea, V. Leiva, and M. Santos Neto. Generalized tobit models: diagnostics and application in econometrics. *Journal of Applied Statistics*, 45:1–23, 12 2016. doi: 10.1080/02664763.2016.1268572.

R. E. Benestad, K. M. Parding, H. B. Erlandsen, and A. Mezghani. A simple equation to study changes in rainfall statistics. *Environmental Research Letters*, 14(8):084017, jul 2019. doi: 10.1088/1748-9326/ab2bb2.

J. Berkson and R. P. Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515, 1952. ISSN 01621459.

V. Bewick, L. Cheek, and J. Ball. Statistics review 12: Survival analysis. *Critical Care*, 8 (5), 2004. doi: 10.1186/cc2955.

H. W. Block and A. Basu. A continuous bivariate exponential extension. *Journal of the American Statistical Association*, 69(348):1031–1037, 1974.

J. W. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):15–53, 1949. ISSN 00359246.

G. B. Bonan. Forests and climate change: Forcings, feedbacks, and the climate benefits of forests. *Science*, 320(5882):1444–1449, 2008.

R. Canales, A. Wilson, J. Pearce-Walker, M. Verhougstraete, and K. Reynolds. Methods for handling left-censored data in quantitative microbial risk assessment. *Applied and Environmental Microbiology*, 84:AEM.01203–18, 08 2018. doi: 10.1128/AEM.01203-18.

V. G. Cancho and H. Bolfarine. Modeling the presence of immunes by using the exponentiated-Weibull model. *Journal of Applied Statistics*, 28(6):659–671, 2001.

B. P. Carlin and T. A. Louis. *Bayes and empirical Bayes methods for data analysis*. Chapman and Hall/CRC, 2010.

J. M. Carrasco, E. M. Ortega, and G. M. Cordeiro. A generalized modified weibull distribution for lifetime modeling. *Computational Statistics & Data Analysis*, 53(2): 450–462, 2008. ISSN 0167-9473. doi: 10.1016/j.csda.2008.08.023. URL https://www.sciencedirect.com/science/article/pii/S0167947308004192.

Z. Chen. A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics & Probability Letters*, 49(2):155–161, 2000.

S. Chib. Bayes inference in the tobit censored regression model. *Journal of Econometrics*, 51 (1):79 – 99, 1992. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(92)90030-U.

S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

C. Chiosi. Fundamentals of stellar evolution theory: Understanding the hrd. *Stellar astrophysics for the local group: VIII Canary Islands Winter School of Astrophysics*, page 1, 1998.

D. Clayton and J. Cuzick. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society: Series A (General)*, 148(2):82–108, 1985.

D. Collett. *Modelling survival data in medical research*. Chapaman and Hall, New York, second edition, 2003.

G. W. Collins. The fundamentals of stellar astrophysics. *New York, WH Freeman and Co., 1989, 512 p.*, 1989.

A. Costello, M. Abbas, A. Allen, S. Ball, S. Bell, R. Bellamy, S. Friel, N. Groce, A. Johnson, M. Kett, M. Lee, C. Levy, M. Maslin, D. McCoy, B. McGuire, H. Montgomery, D. Napier, C. Pagel, J. Patel, J. A. P. de Oliveira, N. Redclift, H. Rees, D. Rogger, J. Scott, J. Stephenson, J. Twigg, J. Wolff, and C. Patterson. Managing the health effects of climate change: Lancet and university college london institute for global health commission. *Lancet (London, England)*, 373(9676):1693—1733, May 2009. ISSN 0140-6736. doi: 10.1016/s0140-6736(09)60935-1. URL https://doi.org/10.1016/S0140-6736(09)60935-1.

C. Cox, H. Chu, M. F. Schneider, and A. Muñoz. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine*, 26(23):4352–4374, 2007. doi: 10.1002/sim.2836.

D. R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220, 1972.

D. R. Cox and D. Oakes. *Analysis of survival data*. Chapman & Hall, London, 1984.

M. J. Crowder. *Multivariate survival analysis and competing risks*. CRC Press, 2012.

R. De Angelis, R. Capocaccia, T. Hakulinen, B. Soderman, and A. Verdecchia. Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in Medicine*, 18(4):441–454, 1999.

R. P. de Oliveira, M. V. de Oliveira Peres, J. A. Achcar, and N. Davarzani. Inference for the trivariate Marshall-Olkin-Weibull distribution in presence of right-censored data. *Chilean Journal of Statistics (ChJS)*, 11(2).

R. P. de Oliveira, J. A. Achcar, D. Peralta, and J. Mazucheli. Discrete and continuous bivariate lifetime models in presence of cure rate: a comparative study under Bayesian approach. *Journal of Applied Statistics*, 46(3):1–19, 2018.

R. P. de Oliveira, A. F. Menezes, J. Mazucheli, and J. A. Achcar. Mixture and nonmixture cure fraction models assuming discrete lifetimes: Application to a pelvic sarcoma dataset. *Biometrical Journal*, 61(4):813–826, 2019.

R. P. de Oliveira, J. A. Achcar, J. Mazucheli, and W. Bertoli. A new class of bivariate Lindley distributions based on stress and shock models and some of their reliability properties. *Reliability Engineering & System Safety*, 211:107528, 2021.

M. F. Desousa. Two essays on birnbaum-saunders regression models for censored data. Master's thesis, Universidade Federal de Goiás, Faculdade de Administração, Ciências Contábeis e Ciências Econômicas (FACE), Programa de Pós-Graduação em Economia (PPE), Goiânia - GO, 2016.

C. A. dos Santos and J. A. Achcar. A bayesian analysis for the block and basu bivariate exponential distribution in the presence of covariates and censored data. *Journal of Applied Statistics*, 38(10):2213–2223, 2011.

F. Downton. Bivariate exponential distributions in reliability theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(3):408–417, 1970.

N. Dzupire, P. Ngare, and L. Odongo. A poisson-gamma model for zero inflated rainfall data. *Journal of Probability and Statistics*, 2018. doi: 10.1155/2018/1012647.

S. Eryilmaz and F. Tank. On reliability analysis of a two-dependent-unit series system with a standby unit. *Applied Mathematics and Computation*, 218(15):7792–7797, 2012.

V. T. Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38(4):1041–1046, 1982.

S. Fatimah, S. F. Che Osmi, M. Abdul Malek, and M. Yusoff. Prediction of ammoniacal nitrogen in river using modified cuckoo search-neural network (cs-nn). *Ecology, Environment and Conservation*, 23:85–88, 11 2017.

L. M. Fernandes. Inferência Bayesiana em modelos discretos com fração de cura. Master's thesis, 2014.

T. R. Fleming and D. Lin. Survival analysis in clinical trials: past developments and future directions. *Biometrics*, 56(4):971–983, 2000.

E. W. Frees. *Regression Modeling with Actuarial and Financial Applications*. International Series on Actuarial Science. Cambridge University Press, 2009. doi: 10.1017/CBO9780511814372.

J. E. Freund. A bivariate extension of the exponential distribution. *Journal of the American Statistical Association*, 56(296):971–977, 1961.

A. E. Gelfand and A. F. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, PAMI-6(6):721–741, 1984. doi: 10.1109/TPAMI.1984.4767596.

J. George, J. Letha, and P. Jairaj. Daily rainfall prediction using generalized linear bivariate model – a case study. *Procedia Technology*, 24:31–38, 2016. ISSN 2212-0173. International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015).

W. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis. CRC Press, 1995. URL https://books.google.de/books?id=TRXrMWY_i2IC.

W. Gilks, S. Richardson, and D. Spieglhalter. Markov chain monte carlo in practice, 520 pp, 1996.

S. R. Giolo and E. A. Colosimo. *Análise de sobrevivência aplicada*. Edgard Blucher, 2006.

E. J. Gumbel. Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292):698–707, 1960.

G. Guo. Event-history analysis for left-truncated data. *Sociological Methodology*, 23: 217–243, 1993. ISSN 00811750, 14679531.

M. Gurvich, A. Dibenedetto, and S. Ranade. A new statistical distribution for characterizing the random strength of brittle materials. *Journal of Materials Science*, 32:2559–2564, 05 1997.

A. G. Hawkes. A bivariate exponential distribution with applications to reliability. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(01):129–131, 1972.

E. Hawkins, P. Ortega, E. Suckling, A. Schurer, G. Hegerl, P. Jones, M. Joshi, T. J. Osborn, V. Masson-Delmotte, J. Mignot, P. Thorne, and G. J. van Oldenborgh. Estimating changes in global temperature since the preindustrial period. *Bulletin of the American Meteorological Society*, 98(9):1841 – 1856, 2017. doi: 10.1175/BAMS-D-16-0007.1.

J. L. Haybittle. A two-parameter model for the survival curve of treated cancer patients. *Journal of the American Statistical Association*, 60(309):16–26, 1965. ISSN 01621459. URL http://www.jstor.org/stable/2283134.

P. Hougaard. A class of multivariate failure time distributions. *Biometrika*, pages 671–678, 1986.

P. Hougaard. *Analysis of multivariate survival data*. Springer Science & Business Media, 2012.

J. G. Ibrahim, M.-H. Chen, and D. Sinha. *Bayesian survival analysis*. Wiley Online Library, 2005.

IPCC. *Summary for policymakers*, page 17. Cambridge University Press, Cambridge, UK, 2007.

IPCC. *Summary for Policymakers*, book section SPM, page 1–30. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013. ISBN ISBN 978-1-107-66182-0. doi: 10.1017/CBO9781107415324.004. URL www.climatechange2013.org.

H. Jacqmin-Gadda, R. Thiébaut, G. Chene, and D. Commenges. Analysis of left-censored longitudinal data with application to viral load in hiv infection. *Biostatistics*, 1 4:355–68, 2000.

J. Job, N. Halsey, R. Boulos, E. Holt, D. Farrell, P. Albrecht, J. Brutus, M. Adrien, J. Andre, E. Chan, and et al. Successful immunization of infants at 6 months of age with high dose Edmonston-Zagreb measles vaccine. Soleil/JHU Project Team. *Pediatr Infect Dis J.*, 10(4):303–311, 1991. doi: 10.1097/00006454-199104000-00008.

R. Kabir, H. Khan, K. Caldwell, and E. Ball. Climate change impact: The experience of the coastal areas of bangladesh affected by cyclones sidr and aila. *Journal of Environmental and Public Health*, 2016. doi: 10.1155/2016/9654753.

D. Kaczan and J. Orgill Meyer. The impact of climate change on migration: a synthesis of recent empirical insights. *Climatic Change*, 158(3):281–300, 2020.

J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley-Interscience, New York, NY, 2nd edition, 2002.

Karl, T. R., Melillo, J. M., and P. and Hassol, S. J. *Global Climate Change Impacts in the United States: A State of Knowledge Report from the U.S. Global Change Research Program*. Cambridge University Press, New York, USA, 2009.

J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, New York, 1997.

S. W. Lagakos and J. S. Williams. Models for censored survival analysis: A cone class of variable-sum models. *Biometrika*, 65(1):181–189, 04 1978. ISSN 0006-3444. doi: 10.1093/biomet/65.1.181.

C. Lai, M. Xie, and D. Murthy. A modified weibull distribution. *IEEE Transactions on Reliability*, 52(1):33–37, 2003. doi: 10.1109/TR.2002.805788.

P. C. Lambert, J. R. Thompson, C. L. Weston, and P. W. Dickman. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, 8 (3):576–594, 2006.

L. Latifoglu. A novel combined model for prediction of daily precipitation data using instantaneous frequency feature and bidirectional long short time memory networks. *Environmental Science and Pollution Research*, 05 2021. doi: 10.21203/rs.3.rs-525276/v1.

J. F. Lawless. *Statistical models and methods for lifetime data*. John Wiley & Sons, Inc., New York, 1982. Wiley Series in Probability and Mathematical Statistics.

E. T. Lee and J. W. Wang. *Statistical methods for survival data analysis*. Wiley Series in Probability and Statistics. ohn Wiley & Sons, Inc., Hoboken, NJ, third edition, 2003.

V. Leiva, M. Barros, G. A. Paula, and M. Galea. Influence diagnostics in log-birnbaum–saunders regression models with censored data. *Computational Statistics & Data Analysis*, 51(12):5694 – 5707, 2007. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2006.09.020.

K.-M. Leung, R. M. Elashoff, and A. A. Afifi. Censoring issues in survival analysis. *Annual Review of Public Health*, 18(1):83–104, 1997. doi: 10.1146/annurev.publhealth.18.1.83.

A. Levermann, P. U. Clark, B. Marzeion, G. A. Milne, D. Pollard, V. Radic, and A. Robinson. The multimillennial sea-level commitment of global warming. *Proceedings of the National Academy of Sciences*, 110(34):13745–13750, 2013. doi: 10.1073/pnas.1219414110.

J. C. Lindsey and L. M. Ryan. Methods for interval-censored data. *Statistics in Medicine*, 17(2):219–238, 1998.

M. Lineman, Y. Do, J. Y. Kim, and G.-J. Joo. Talking about climate change and global warming. *PLOS ONE*, 10(9):1–12, 09 2015.

D. B. Lobell and M. B. Burke. On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, 150(11):1443–1452, 2010. ISSN 0168-1923.

J. Long, J. Long, and J. Freese. *Regression Models for Categorical and Limited Dependent Variables*. Advanced Quantitative Techniques in the Social Sciences. SAGE Publications, 1997. ISBN 9780803973749.

A. M. López, J. T. Cuevas, and C. Gutiérrez. Niveles de tiroglobulina previo a la ablacón y persistencia / recurrencia precoz del cáncer diferenciado de tiroides. *Revista Ciencias de la Salud*, 12(1):9–21, 2014. ISSN 2145-4507.

W. Lu. Efficient estimation for an accelerated failure time model with a cure fraction. *Statistica Sinica*, 20:661–674, 2010.

H. Lynn. Maximum likelihood inference for left-censored hiv rna data. *Statistics in Medicine*, 20:33–45, 02 2001. doi: 10.1002/1097-0258(20010115)20:13.0.CO;2-O.

R. A. Maller and X. Zhou. *Survival Analysis With Long-Term Survivors*. Wiley New York, 1996.

A. W. Marshall and I. Olkin. A multivariate exponential distribution. *Journal of the American Statistical Association*, 62(317):30–44, 1967.

A. W. Marshall and I. Olkin. A family of bivariate distributions generated by the bivariate Bernoulli distribution. *Journal of the American Statistical Association*, 80(390):332–338, 1985.

E. Z. Martinez, J. A. Achcar, A. A. A. Jácome, and J. S. Santos. Mixture and non-mixture cure fraction models based on the generalized modified weibull distribution with an application to gastric cancer data. *Computer methods and programs in biomedicine*, 112 (3):343—355, December 2013. ISSN 0169-2607. doi: 10.1016/j.cmpb.2013.07.021.

G. Martínez-Flórez, H. Bolfarine, and H. W. Gómez. The alpha-power tobit model. *Communications in Statistics - Theory and Methods*, 42(4):633–643, 2013. doi: 10.1080/03610926.2011.630770.

T. Matthews. Humid heat and climate change. *Progress in Physical Geography: Earth and Environment*, 42(3):391–405, 2018.

C. McGilchrist and C. Aisbett. Regression with frailty in survival analysis. *Biometrics*, 47 (2):461–466, 1991.

G. J. McLachlan and D. Peel. *Finite mixture models / Geoffrey McLachlan, David Peel*. Wiley New York ; Chichester, 2000.

S. Mitra and D. Kundu. Analysis of left censored data from the generalized exponential distribution. *Journal of Statistical Computation and Simulation*, 78(7):669–679, 2008. doi: 10.1080/00949650701344158.

K. Mosler. Mixture models in econometric duration analysis. *Applied Stochastic Models in Business and Industry*, 19(2):91–104, 2003. doi: 10.1002/asmb.489.

L. H. Moulton and N. A. Halsey. A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics*, 51(4):1570–1578, 1995. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2533289.

G. Mudholkar and D. Srivastava. Exponentiated weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability*, 42(2):299–302, 1993. doi: 10.1109/24.229504.

K. Muralidharan and P. Lathika. Analysis of instantaneous and early failures in weibull distribution. *Metrika*, 64:305–316, 02 2006. doi: 10.1007/s00184-006-0050-2.

S. Nadarajah and S. Kotz. On some recent modifications of weibull distribution. *Reliability, IEEE Transactions on*, 54:561 – 562, 01 2006. doi: 10.1109/TR.2005.858811.

R. B. Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

M. Nikulin and F. Haghighi. A chi-squared test for the generalized power weibull family for the head-and-neck cancer censored data. *Journal of Mathematical Sciences*, 133:1333–1341, 03 2006. doi: 10.1007/s10958-006-0043-8.

D. Oakes. Semiparametric inference in a model for association in bivanate survival data. *Biometrika*, 73(2):353–361, 1986.

D. Oakes. Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84(406):487–493, 1989.

R. Oliveira, J. Mazucheli, and J. Achcar. A generalization of basu-dhar's bivariate geometric distribution to the trivariate case. *Communications in Statistics - Simulation and Computation*, 2019. doi: 10.1080/03610918.2019.1643881.

E. M. M. Ortega, G. M. Cordeiro, and M. W. Kattan. The log-beta weibull regression model with application to predict recurrence of prostate cancer. *Statistical Papers*, 154:113 – 132, 2013.

M. Othus, B. Barlogie, M. L. LeBlanc, and J. J. Crowley. Cure models as a useful statistical tool for analyzing survival. *Clinical Cancer Research*, 18(14):3731–3736, 2012.

M. Pal, M. M. Ali, and J. Woo. Exponentiated weibull distribution. *Statistica*, 66(2): 139–147, Jan. 2006.

S. Petterson, R. Grondahl-Rosado, V. Nilsen, M. Myrmel, and L. Robertson. Variability in the recovery of a virus concentration procedure in water: Implications for qmra. *Water research*, 87:79–86, 09 2015. doi: 10.1016/j.watres.2015.09.006.

H. Pham and C.-D. Lai. On recent generalizations of the Weibull distribution. *Transactions on Reliability*, 56(3):454–458, 2007.

E. Poloczanska, C. Brown, W. Sydeman, W. Kiessling, D. Schoeman, P. Moore, K. Brander, J. Bruno, L. Buckley, and M. Burrows. Global imprint of climate change on marine life. *Nature Climate Change*, 3, 08 2013. doi: 10.1038/nclimate1958.

R. Pouillot, J. M. Van Doren, J. Woods, D. Plante, M. Smith, G. Goblick, C. Roberts, A. Locas, W. Hajen, J. Stobo, J. White, J. Holtzman, E. Buenaventura, W. Burkhardt, A. Catford, R. Edwards, A. DePaola, and K. R. Calci. Meta-analysis of the reduction of norovirus and male-specific coliphage concentrations in wastewater treatment plants. *Applied and Environmental Microbiology*, 81(14):4669–4681, 2015. ISSN 0099-2240. doi: 10.1128/AEM.00509-15. URL https://aem.asm.org/content/81/14/4669.

D. L. Price and A. K. Manatunga. Modelling survival data with a cured fraction using frailty models. *Statistics in Medicine*, 20(9-10):1515–1527, 2001.

R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL http://www.R-project.org. ISBN 3-900051-07-0.

S. Rahmstorf, A. Cazenave, J. A. Church, J. E. Hansen, R. F. Keeling, D. E. Parker, and R. C. J. Somerville. Recent climate observations compared to projections. *Science*, 316 (5825):709–709, 2007. doi: 10.1126/science.1136843. URL https://www.science.org/doi/abs/10.1126/science.1136843.

M. Rausand and A. Hoyland. *System reliability theory: models, statistical methods, and applications*. John Wiley & Sons, 2004. ISBN ISBN 0-471-47133-X.

G. R. Richards. Change in global temperature: A statistical analysis. *Journal of Climate*, 6(3):546 – 559, 1993.

J. L. Romeu. Understanding series and parallel systems reliability. *Selected Topics in Assurance Related Technologies (START), Department of Defense Reliability Analysis Center (DoD RAC)*, 11(5), 2004.

N. Santos, G. Israelian, R. G. López, M. Mayor, R. Rebolo, S. Randich, A. Ecuvillon, and C. D. Cerdeña. Are beryllium abundances anomalous in stars with giant planets? *Astronomy & Astrophysics*, 427(3):1085–1096, 2004.

O. Serdeczny, S. Adams, F. Baarsch, D. Coumou, A. Robinson, W. Hare, M. Schaeffer, M. Perrette, and J. Reinhardt. Climate change impacts in sub-saharan africa: from physical changes to their social repercussions. *Regional Environmental Change*, 17(6): 1585–1600, 2017. ISSN 1436-3798. doi: 10.1007/s10113-015-0910-2.

J. H. Shih. Models and analysis for multivariate failure time data. 1992.

M. Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. Winbugs user manual, 2003.

M. Springmann, D. Mason-D'Croz, S. Robinson, T. Garnett, C. Godfray, D. Gollin, M. Rayner, P. Ballon, and P. Scarborough. Global and regional health effects of future food production under climate change: A modelling study. *The Lancet*, 387, 03 2016. doi: 10.1016/S0140-6736(15)01156-3.

V. Sreeja and P. Sankaran. *Regression models for bivariate survival data.* PhD thesis, Cochin University of Science and Technology, 2008.

R. D. Stern and R. Coe. A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society. Series A (General)*, 147(1):1–34, 1984.

C. A. Struthers and V. T. Farewell. A mixture model for time to aids data with left truncation and an uncertain origin. *Biometrika*, 76(4):814–817, 12 1989. doi: 10.1093/biomet/76.4.814.

T. L. Thorarinsdottir and T. Gneiting. Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society Series A*, 173(2):371–388, 2010.

D. Titterington, P. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. Applied section. Wiley, 1985. ISBN 9780471907633. URL https://books.google.com.br/books?id=hZ0QAQAAIAAJ.

J. Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26 (1):24–36, 1958.

P. K. Trivedi and D. M. Zimmer. *Copula modeling: an introduction for practitioners*. Now Publishers Inc, 2007.

A. Tsodikov. A proportional hazards model taking account of long-term survivors. *Biometrics*, 54(4):1508–1516, 1998. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2533675.

A. Tsodikov, J. Ibrahim, and A. Yakovlev. Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, 98(464):1063–1078, 2003.

M. Turner, W. Calder, G. Cumming, T. Hughes, A. Jentsch, S. LaDeau, T. Lenton, B. Shuman, M. Turetsky, Z. Ratajczak, J. Williams, A. Williams, and S. Carpenter. Climate change, ecosystems and abrupt change: Science priorities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375:20190105, 03 2020. doi: 10.1098/rstb.2019.0105.

J. W. Vaupel. How change in age-specific mortality affects life expectancy. *Population Studies*, 40(1):147–157, 1986.

G. Vergara, J. Rose, and K. Gin. Risk assessment of noroviruses and human adenoviruses in recreational surface waters. *Water Research*, 103(15):276–282, 2016. doi: 10.1016/j.watres.2016.07.048.

C. Villegas, G. Paula, and V. Leiva. Birnbaum-saunders mixed models for censored reliability data analysis. *IEEE Transactions on Reliability*, 60:748–758, 2011. doi: 10.1109/TR.2011.2170251.

W. Weibull. *A Statistical Theory of the Strength of Materials*. Generalstabens litografiska anstalts forlag, 1939.

W. Weibull. Wide applicability. *Journal of Applied Mechanics*, 103(730):293–297, 1951.

D. Wilks. Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, 210(1):178–191, 1998. ISSN 0022-1694.

M. Xie, Y. Tang, and T. N. Goh. A modified Weibull extension with bathtub-shaped failure rate function. *Reliability Engineering & System Safety*, 76(3):279–285, 2002.

M.-H. Yeo, H.-L. Nguyen, and V.-T.-V. Nguyen. Statistical tool to modeling of a daily precipitation process in the context of climate change. *Journal of Water and Climate Change*, 12, 11 2019. doi: 10.2166/wcc.2019.403.

G. Yin and J. G. Ibrahim. Cure rate models: a unified approach. *Canadian Journal of Statistics*, 33(4):559–570, 2005.

B. Yu, R. C. Tiwari, K. A. Cronin, and E. J. Feuer. Cure fraction estimation from the mixture cure models for grouped survival data. *Statistics in Medicine*, 23(11):1733–1747, 2004.

C. Zhao, B. Liu, S. Piao, X. Wang, D. B. Lobell, Y. Huang, M. Huang, Y. Yao, S. Bassu, P. Ciais, J.-L. Durand, J. Elliott, F. Ewert, I. A. Janssens, T. Li, E. Lin, Q. Liu, P. Martre, C. Muller, S. Peng, J. Penuelas, A. C. Ruane, D. Wallach, T. Wang, D. Wu, Z. Liu, Y. Zhu, Z. Zhu, and S. Asseng. Temperature increase reduces global yields of major crops in four independent estimates. *Proceedings of the National Academy of Sciences*, 114(35):9326–9331, 2017. doi: 10.1073/pnas.1701762114.

L. Zhiying and H. Fang. Impacts of climate change on water erosion: A review. *Earth-Science Reviews*, 163:94–117, 2016.

Appendix- R codes

Appendix 1

```
# Tobit Weibull model no covariates

model.jags.weib.tobit  <-  function()
{
c    <-  0.01
for(i in 1:n)
    {
    phi[i]  <--(L[i])
    zeros[i]~dpois(phi[i])

    pdf[i]   <-  log(alpha)-alpha*log(beta)+(alpha)*log(x[i])-(x[i]/beta)^alpha
    surv[i]  <-  -(c/beta)^alpha

    L[i]  <-  (1 - delta[i]) * log(p) + (delta[i]) * (log(1-p) +
                  pdf[i] - surv[i])
    }

    # Priors
    alpha~dgamma(0.01,0.01)
    beta~dgamma(0.01,0.01)
    p~dbeta(7,3)
    }
```

```
# Tobit Weibull Model with covariates

model.jags.weib  <-function()
{
c         <-  0.01
for(i in 1:n)
    {
    phi[i] <--log(L[i])
    zeros[i]~dpois(phi[i])

    beta[i]     <- exp(omega0 + omega1 * cov1[i] + omega2 * cov2[i] +
                 + omega3 * cov3[i] + omega4 * cov4[i] + omega5 * cov5[i] +
                 + omega6 * cov6[i] + omega7 * cov7[i] + omega8 * cov8[i] +
                 + omega9 * cov9[i] + omega10 * cov10[i]+omega11 * cov11[i] +
                 + omega12 * cov12[i] + omega13 * cov13[i])

    logit(p[i]) <- psi0 +  psi1 * cov1[i] + psi2 * cov2[i] +
                 + psi3 * cov3[i] + psi4 *cov4[i] + psi5 * cov5[i] +
                 + psi6 * cov6[i] + psi7 * cov7[i] + psi8 * cov8[i] +
                 + psi9 * cov9[i] + psi10 * cov10[i] + psi11 * cov11[i] +
                 + psi12 * cov12[i] + psi13 * cov13[i]

    L1[i]    <-  (1 - delta[i]) * log(p[i]) + (delta[i]) * (log(1-p[i]) +
                   log(alpha)-alpha*log(beta[i])+(alpha)*log(x[i]) +
                   - (x[i]/beta[i])^alpha + (c/beta[i])^alpha)
    L[i]    <-  exp(L1[i])
    }

# Prioris
alpha~dgamma(0.01,0.01)
omega0~dnorm(0,1)
omega1~dnorm(0,1)
omega2~dnorm(0,1)
omega3~dnorm(0,1)
omega4~dnorm(0,1)
omega5~dnorm(0,1)
```

```
omega6~dnorm(0,1)
omega7~dnorm(0,1)
omega8~dnorm(0,1)
omega9~dnorm(0,1)
omega10~dnorm(0,1)
omega11~dnorm(0,1)
omega12~dnorm(0,1)
omega13~dnorm(0,1)
psi0~dbeta(7,3)
psi1~dnorm(0,1)
psi2~dnorm(0,1)
psi3~dnorm(0,1)
psi4~dnorm(0,1)
psi5~dnorm(0,1)
psi6~dnorm(0,1)
psi7~dnorm(0,1)
psi8~dnorm(0,1)
psi9~dnorm(0,1)
psi10~dnorm(0,1)
psi11~dnorm(0,1)
psi12~dnorm(0,1)
psi13~dnorm(0,1)
}
```

Appendix 2.1 - Stellar Dataset

The columns of the dataset in Table 20 are: star name; Type = 1 indicates planet-hosting stars and Type = 2 is the control sample; Teff (in degrees Kelvin) is the stellar surface temperature; log N(Be), log of the abundance of beryllium scaled to the Sun's abundance (i.e., the Sun has log N(Be) = 0.0); log N(Li), log of the abundance of lithium scaled to the Sun's abundance. The indicator variables of left-censoring are given by $\delta_j = 1$ if T is a complete observation and $\delta_j = 0$ if T is a left censored observation, j = 1 (Be) and j = 2 (Li).

Table 20 – Stellar astronomy data set.

Row	Star	Type	Teff	δ_1	log[N(Be)]	δ_2	log[N(Li)]
1	HD-6434	1	5835	1	1.08	0	0.80
2	HD-9826	1	6212	1	1.05	1	2.55
3	HD-10647	1	6143	1	1.19	1	2.80
4	HD-10697	1	5641	1	1.31	1	1.96
5	HD-12661	1	5702	1	1.13	0	0.98
6	HD-13445	1	5613	0	0.40	0	-0.12
7	HD-16141	1	5801	1	1.17	1	1.11
8	HD-17051	1	6252	1	1.03	1	2.66
9	HD-19994	1	6109	1	0.93	1	1.99
10	HD-22049	1	5073	1	0.77	0	0.25
11	HD-27442	1	4825	0	0.30	0	-0.47
12	HD-38529	1	5674	0	-0.10	0	0.61
13	HD-46375	1	5268	0	0.80	0	-0.02
14	HD-52265	1	6103	1	1.25	1	2.88
15	HD-75289	1	6143	1	1.36	1	2.85
16	HD-82943	1	6016	1	1.27	1	2.51
17	HD-92799	1	5821	1	1.19	1	1.34
18	HD-95128	1	5924	1	1.23	1	1.83
19	HD-108147	1	6248	1	0.99	1	2.33
20	HD-114762	1	5884	1	0.82	1	2.20
21	HD-117176	1	5560	1	0.86	1	1.88
22	HD-121504	1	6075	1	1.33	1	2.65
23	HD-130322	1	5392	1	0.95	0	0.13
24	HD-134987	1	5776	1	1.22	0	0.74
25	HD-143761	1	5853	1	1.11	1	1.46
26	HD-145675	1	5311	0	0.65	0	0.03
27	HD-169830	1	6299	0	-0.40	0	1.16
28	HD-179949	1	6260	1	1.08	1	2.65
29	HD-187123	1	5845	1	1.08	1	1.21
30	HD-192263	1	4947	0	0.90	0	-0.39

cont...

Row	Star	Type	Teff	δ_1	log[N(Be)]	δ_2	log[N(Li)]
31	HD-195019	1	5842	1	1.15	1	1.47
32	HD-202206	1	5752	1	1.04	1	1.04
33	HD-209458	1	6117	1	1.24	1	2.70
34	HD-210277	1	5532	1	0.91	0	0.30
35	HD-217014	1	5804	1	1.02	1	1.30
36	HD-217107	1	5646	1	0.96	0	0.40
37	HD-222582	1	5843	1	1.14	0	0.59
38	HD-870	2	5447	1	0.80	0	0.20
39	HD-1461	2	5768	1	1.14	0	0.51
40	HD-1581	2	5956	1	1.15	1	2.37
41	HD-3823	2	5948	1	1.02	1	2.41
42	HD-4391	2	5878	1	0.75	0	1.09
43	HD-7570	2	6140	1	1.17	1	2.91
44	HD-10700	2	5344	1	0.83	0	0.41
45	HD-14412	2	5368	1	0.80	0	0.44
46	HD-20010	2	6275	1	1.01	1	2.13
47	HD-20766	2	5733	0	-0.09	0	0.97
48	HD-20794	2	5444	1	0.91	0	0.52
49	HD-20807	2	5843	1	0.36	0	1.07
50	HD-23249	2	5074	0	0.15	1	1.24
51	HD-23484	2	5176	0	0.70	0	0.40
52	HD-26965A	2	5126	1	0.76	0	0.17
53	HD-30495	2	5768	1	1.16	1	2.44
54	HD-36435	2	5479	1	0.99	1	1.67
55	HD-38858	2	5752	1	1.02	1	1.64
56	HD-43162	2	5633	1	1.08	1	2.34
57	HD-43834	2	5594	1	0.94	1	2.30
58	HD-69830	2	5410	1	0.79	0	0.47
59	HD-72673	2	5242	1	0.70	0	0.48
60	HD-74576	2	5000	1	0.70	1	1.72
61	HD-76151	2	5803	1	1.02	1	1.88
62	HD-85117	2	6167	1	1.11	1	2.64
63	HD-189567	2	5765	1	1.06	0	0.82
64	HD-192310	2	5069	0	0.60	0	0.20
65	HD-211415	2	5890	1	1.12	1	1.92
66	HD-222335	2	5260	1	0.66	0	0.31

Appendix 2.2 - OpenBugs Codes

```
Model 1

model
{
for (i in 1:N)
{
    zeros[i] <- 0
    dummy[i] <- 0
    dummy[i] ~ dloglik(logLike[i])
    a1[i]<- Be[i]/beta1[i]
    a2[i]<- Li[i]/beta2[i]

    logLike[i]    <- (1-delta.Be[i])*log(1-exp(-pow(a1[i], alpha1))) +
                     delta.Be[i]*(log(alpha1)-alpha1*log(beta1[i]) +
                     (alpha1-1)*log(Be[i])-pow(a1[i],alpha1)) +
                     (1-delta.Li[i])*log(1-exp(-pow(a2[i],alpha2))) +
                     delta.Li[i]*(log(alpha2)-alpha2*log(beta2[i]) +
                     (alpha2-1)*log(Li[i])-pow(a2[i],alpha2))

    log(beta1[i]) <- gamma10 + gamma11*type[i] + gamma12*log(teff[i]) + w[i]
    log(beta2[i]) <- gamma20 + gamma21*type[i] + gamma22*log(teff[i]) =
                    + gamma23*pow(log(teff[i]),2)+w[i]
    w[i] ~ dnorm(0,tau)
}
alpha1 ~ dunif(0,10)
alpha2 ~ dunif(0,10)
gamma10 ~ dnorm(0,0.01)
gamma11 ~ dnorm(0,1)
gamma12 ~ dnorm(0,1)
gamma20 ~ dnorm(0,0.01)
gamma21 ~ dnorm(0,1)
gamma22 ~ dnorm(0,1)
gamma23 ~ dnorm(0,10)
tau ~ dunif(0,200)
}
```

```
Model 2

model
{
for (i in 1:N)
    {
    dummy[i] <- 0
    dummy[i] ~ dloglik(logLike[i])

    a11[i] <- (Be[i]/beta1[i])
    a12[i] <- pow(a11[i],alpha1)
    b11[i] <- (c1[i]/beta1[i])
    b12[i] <- pow(b11[i],alpha1)
    a21[i] <- (Li[i]/beta2[i])
    a22[i] <- pow(a21[i],alpha2)
    b21[i] <- (c2[i]/beta2[i])
    b22[i] <- pow(b21[i],alpha2)

    log(beta1[i]) <- gamma10+gamma11*type[i]+gamma12*log(teff[i])+w[i]
    log(beta2[i]) <- gamma20+gamma21*type[i]+gamma22*log(teff[i])
    +gamma23*pow(log(teff[i]),2)+w[i]

    logit(p1[i])  <- tau10 + tau11*type[i] + tau12*log(teff[i]) + w[i]
    logit(p2[i])  <- tau20 + tau21*type[i] + tau22*log(teff[i]) +
    w[i] ~ dnorm(0,tau)

    logLike[i]    <- (1-delta.Be[i])*log(p1[i]) + (delta.Be[i])*(log(1-p1[i])+
          log(alpha1) + (alpha1-1)*log(Be[i]) - alpha1*log(beta1[i]) -
          a12[i]+b12[i]) +  (1-delta.Li[i])*log(p2[i]) +
          delta.Li[i])*(log(1-p2[i]) + log(alpha2) + (alpha2-1)*log(Li[i]) -
          alpha2*log(beta2[i]) - a22[i]+b22[i])

    log(v1[i])  <- log(1-p1[i])  + log(alpha1) + (alpha1-1)*log(Be[i]) -
    alpha1*log(beta1[i]) - a12[i] + b12[i]
    log(v2[i])  <- log(1-p2[i]) + log(alpha2) + (alpha2-1)*log(Li[i]) -
    alpha2*log(beta2[i]) - a22[i]+b22[i]
```

```
    theta1[i]   <- v1[i]/(p1[i]+v1[i])
    theta2[i]   <- v2[i]/(p2[i]+v2[i])
    delta.Be[i] ~ dbern(theta1[i])
    delta.Li[i] ~ dbern(theta2[i])
    }

alpha1 ~ dgamma(1,1)
alpha2 ~ dgamma(1,1)
gamma10 ~ dnorm(0,1)
gamma11 ~ dnorm(0,100)
gamma12 ~ dnorm(0,100)
gamma20 ~ dnorm(0,1)
gamma21 ~ dnorm(0,100)
gamma22 ~ dnorm(0,100)
gamma23 ~ dnorm(0,100)
tau10 ~ dnorm(0,1)
tau11 ~ dnorm(0,100)
tau12 ~ dnorm(0,100)
tau20 ~ dnorm(0,1)
tau21 ~ dnorm(0,100)
tau22 ~ dnorm(0,100)
tau23 ~ dnorm(0,100)
tau ~ dgamma(0,100)
}
```

Appendix 3.1 - Thyroid Cancer Data

The data set consists of 91 patients from a descriptive study to assess the relationship between initial thyroglobulin levels and the presence of cancer recurrence one year after treatment. The variables in the data set are: each patient's thyroglobulin level before starting iodine therapy (T_1); the thyroglobulin measurement approximately 6 months after the last session (T_2) and (T_3) the thyroglobulin level measurement approximately one year after the last therapy session. The information on thyroglobulin levels is censored, since the measurement instrument does not detect values below 0.1. Other covariates: sex (male=0; female =1); size measurements in millimeters, dosis and persistence measurements as nanograms per milliliter (if $TG < 2ng/ml = 0$ and $TG \geq 2ng/ml = 1$).

row	age	sex	size	dosis131	persist	Tg1	delta1	Tg2	delta2	Tg3	delta3
1	53	1	1	150	0	0.4	1	0.1	1	0.1	0
2	60	1	1	150	0	0.6	1	0.1	0	0.1	0
3	43	1	3	173	0	13	1	0.9	1	0.6	1
4	26	1	4	150	0	5.6	1	2.4	1	0.3	1
5	50	1	4	150	0	23.6	1	0.43	1	0.1	0
6	57	1	5	150	0	3.5	1	0.1	0	0.1	0
7	50	1	6	119	0	0.5	1	0.1	0	0.1	0
8	48	1	6	150	0	1.5	1	0.2	1	0.2	1
9	39	1	7	150	0	3.3	1	0.3	1	0.3	1
10	43	1	9	150	0	1	1	0.1	1	0.1	0
11	49	1	9	150	0	8.4	1	1.2	1	0.4	1
12	57	1	10	170	0	0.6	1	0.5	1	0.1	1
13	50	1	10	156	1	2.7	1	0.1	1	2.1	1
14	52	1	10	150	0	11	1	0.1	1	0.1	0
15	46	1	10	152	0	184	1	0.7	1	0.2	1
16	65	1	11	151	0	0.2	1	0.1	0	0.1	0
17	45	1	11	157	0	0.8	1	0.2	1	0.2	1
18	51	1	11	153	0	1.4	1	0.1	1	0.1	1
19	50	1	11	153	0	1.4	1	0.1	1	0.1	1
20	56	1	11	150	0	23.8	1	0.1	0	0.1	0
21	66	1	12	110	0	0.1	1	0.2	1	0.1	1
22	55	1	12	178	0	0.2	1	0.1	0	0.1	0
23	64	1	12	154	0	0.3	1	0.2	1	0.1	0
24	35	0	12	150	0	1.6	1	0.1	0	0.1	0
25	64	1	12	155	0	2.4	1	16.6	1	0.1	0
26	42	1	12	160	0	3.1	1	0.54	1	0.1	0
27	52	0	12	164	1	24	1	4	1	2.2	1
28	62	1	12	154	0	27	1	1.4	1	0.6	1
29	21	1	13	112	0	0.1	1	0.3	1	0.84	1
30	28	1	13	150	0	0.1	1	0.1	0	0.1	0
31	52	1	13	150	0	0.4	1	0.1	0	0.1	0
32	27	1	13	125	0	1.2	1	0.2	1	0.2	1
33	63	1	14	155	0	3.3	1	0.6	1	0.2	1
34	42	1	14	104	0	11.8	1	0.3	1	0.1	1
35	54	1	15	173	0	0.1	1	0.1	1	0.1	1
36	31	1	15	50	0	0.7	1	0.1	0	0.2	1
37	52	1	15	150	0	1.4	1	0.2	1	0.1	1
38	49	1	15	160	0	2.1	1	0.1	0	0.1	0
39	71	1	15	220	0	2.6	1	0.4	1	0.4	1
40	53	1	15	150	0	6.9	1	0.1	0	0.1	0
41	72	1	15	150	0	139	1	21	1	0.1	0

cont...

row	age	sex	size	dosis131	persist	Tg1	delta1	Tg2	delta2	Tg3	delta3
42	27	1	17	152	0	0.6	1	0.2	1	0.2	1
43	41	1	17	150	0	5.1	1	0.1	1	0.2	1
44	75	1	17	130	1	275	1	20	1	55	1
45	31	0	17	150	1	3000	1	89	1	142	1
46	46	1	18	170	0	0.9	1	0.1	0	0.1	0
47	31	0	18	150	0	3.5	1	0.2	1	0.1	1
48	56	0	18	143	1	10	1	0.6	1	2.6	1
49	24	1	19	153	0	8.2	1	0.1	1	0.2	1
50	47	1	19	150	0	11	1	0.2	1	0.1	0
51	26	0	19	165	0	11.5	1	0.1	1	0.1	0
52	81	0	19	200	1	2474	1	462	1	586	1
53	44	1	20	157	0	0.5	1	0.1	0	0.1	0
54	65	1	20	150	0	0.5	1	0.4	1	0.49	1
55	47	1	20	161	0	2.5	1	0.2	1	0.1	0
56	59	1	20	151	0	3.2	1	0.1	0	0.1	0
57	30	0	20	50	0	3.4	1	0.1	0	0.1	0
58	36	0	20	163	0	5	1	0.1	0	0.1	0
59	26	0	20	152	0	5.8	1	0.49	1	0.1	0
60	55	1	20	170	1	10.9	1	25	1	19.9	1
61	55	0	20	150	0	18	1	0.8	1	0.6	1
62	54	1	22	150	0	304	1	0.1	0	0.1	0
63	64	1	24	150	1	26.1	1	1.5	1	16.7	1
64	39	0	24	168	1	562	1	19.4	1	8.9	1
65	39	0	24	168	1	562	1	19	1	18.9	1
66	48	1	25	150	0	9.7	1	1.6	1	1.4	1
67	56	1	25	155	1	82.5	1	4.8	1	5.2	1
68	42	0	26	150	0	16.5	1	0.64	1	0.13	1
69	33	1	27	150	0	2.5	1	0.6	1	0.1	0
70	41	1	28	150	0	0.3	1	1.6	1	0.27	1
71	23	1	28	152	1	24.8	1	1.6	1	7.2	1
72	40	1	28	154	1	24.9	1	2.4	1	8.1	1
73	41	1	30	150	0	2.7	1	0.1	0	0.1	0
74	54	1	30	154	0	24	1	0.66	1	0.62	1
75	12	1	30	152	1	59	1	8.9	1	11	1
76	64	1	30	153	1	220.2	1	12	1	3.6	1
77	48	1	32	171	0	7.5	1	0.2	1	0.1	1
78	52	1	35	172	0	5.4	1	0.1	0	0.1	0
79	86	1	36	157	1	44.6	1	13	1	6.8	1
80	23	0	38	150	0	0.3	1	0.1	1	0.1	1
81	59	1	42	157	1	123	1	2.6	1	2.6	1
82	50	1	45	151	0	0.1	1	0.37	1	0.1	1
83	52	1	46	150	0	3.9	1	0.42	1	0.1	1
84	29	1	50	150	0	12	1	1.85	1	1.46	1
85	51	1	50	151	0	15.5	1	2.7	1	0.1	0
86	46	1	55	156	0	19	1	0.5	1	0.6	1
87	18	1	55	150	1	112	1	22.6	1	24	1
88	44	1	57	150	0	18.2	1	0.1	0	0.1	0
89	50	0	60	150	0	3.2	1	0.6	1	0.3	1
90	65	0	62	150	1	672	1	101	1	220	1
91	60	1	85	172	1	281	1	33	1	14.3	1

Appendix 3.2 - R Codes

```
# Tobit-Weibull Bivariate - Thyroid cancer data

model.jags   <- function()
{
for (i in 1:n)
    {
    phi[i]       <--log(logLike[i])
    zeros[i]~dpois(phi[i])

    a11[i]        <- (t1[i]/beta1[i])
    a12[i]        <- pow(a11[i],alpha1)
    b11[i]        <- (c1[i]/beta1[i])
    b12[i]        <- pow(b11[i],alpha1)
    a21[i]        <- (t2[i]/beta2[i])
    a22[i]        <- pow(a21[i],alpha2)
    b21[i]        <- (c2[i]/beta2[i])
    b22[i]        <- pow(b21[i],alpha2)

    log(beta1[i]) <- gamma10 + gamma11*cov1[i] + gamma12*cov2[i] +
                     + gamma13*cov3[i] + w[i]
    log(beta2[i]) <- gamma20 + gamma21*cov1[i] + gamma22*cov2[i] +
                     + gamma23*cov3[i] + w[i]
    logit(p1[i])  <- tau10 + tau11*cov1[i] + tau12*cov2[i] +
                     + tau13*cov3[i] + w[i]
    logit(p2[i])  <- tau20 + tau21*cov1[i] + tau22*cov2[i] +
                     + tau23*cov3[i] + w[i]

    w[i] ~ dnorm(0,tau)

    logLike[i]   <- exp((1 - delta1[i]) * log(p1[i]) +
                    (delta1[i]) * (log(1-p1[i]) +
                    log(alpha1) + (alpha1-1) * log(t1[i]) -
                    alpha1 * log(beta1[i]) - a12[i] + b12[i]) +
                    (1 - delta2[i]) * log(p2[i]) +
                    (delta2[i]) * (log(1-p2[i]) +
                    log(alpha2) + (alpha2-1) * log(t2[i]) -
                    alpha2 * log(beta2[i]) - a22[i] + b22[i]))
  }
# Priors
alpha1 ~ dgamma(0.001,0.001)
alpha2 ~ dgamma(0.001,0.001)
```

```
gamma10 ~ dnorm(0,100)
gamma11 ~ dnorm(0,100)
gamma12 ~ dnorm(0,100)
gamma13 ~ dnorm(0,100)
gamma20 ~ dnorm(0,100)
gamma21 ~ dnorm(0,100)
gamma22 ~ dnorm(0,100)
gamma23 ~ dnorm(0,100)
tau10 ~ dnorm(0,100)
tau11 ~ dnorm(0,100)
tau12 ~ dnorm(0,100)
tau13 ~ dnorm(0,100)
tau20 ~ dnorm(0,100)
tau21 ~ dnorm(0,100)
tau22 ~ dnorm(0,100)
tau23 ~ dnorm(0,100)
tau ~ dgamma(1,100)
}

## Tobit-Weibull model - Vaccine data ##

model.jags.weib     <-  function()
{
    c               <-  0.01
    for(i in 1:n)
    {
        phi[i]     <--(L[i])
        zeros[i]~dpois(phi[i])

        beta[i]     <-  exp(gamma0 + gamma1*x1[i] + gamma2*x2[i] + gamma3*x3[i])

        logit(p[i]) <-  psi0 + psi1*x1[i] + psi2*x2[i] + psi3*x3[i]

        pdf[i]      <- (alpha/beta[i]^alpha) * ((x[i])^(alpha-1))*exp(-(x[i]/beta[i])^alpha)
        surv[i]     <-  exp(-c/beta[i])^alpha

        L[i]        <-  (1 - delta[i]) * log(p[i]) + (delta[i]) * (log(1-p[i])
                        + log(pdf[i]) - log(surv[i]))

        a1[i]       <-  (x[i]/beta[i])^alpha
        a2[2]       <-  (c/beta[i])^alpha
        cdfwei[i]   <-  1 - ((1-p[i]) * exp(-a1[i])/exp(-a2[i]))
        Res[i]      <- -log(cdfwei[i])
   }
# Priors
```

```
alpha~dgamma(1,2)
gamma0~dnorm(0,1)
gamma1~dnorm(0,1)
gamma2~dnorm(0,1)
gamma3~dnorm(0,1)
psi0~dnorm(0,1)
psi1~dnorm(0,1)
psi2~dnorm(0,1)
psi3~dnorm(0,1)
}
```

Appendix 4 - R Codes

```
## Weibull Mixture Model - No covariates ##

## Precipitation

model.jags.weib <-  function()
{
  c    <-  0.01
    for(i in 1:n)
  {
    phi[i]    <--(L[i])
    zeros[i]~dpois(phi[i])

    pdf[i]     <-  log(alpha)-alpha*log(beta)+
                   (alpha-1)*log(x[i])-(x[i]/beta)^alpha
    surv[i]    <-  -(c/beta)^alpha

    L[i]    <-  (1 - delta[i]) * log(p) +
                   (delta[i]) * (log(1-p) + pdf[i]- surv[i])
     }

  ## Priors
  alpha~dunif(0.1,2)
  beta~dunif(0.1,100)
  p~dunif(0,1)
}


## Pressure

model.jags.m2 <-  function()
{
  for (i in 1:n)
  {
    x[i] ~ dnorm(mu,tau)
  }
   variance  <- 1/tau
   desvpad   <- sqrt(variance)

  mu ~ dnorm(900,0.001)
  tau~ dunif(0,1)
}
```

```
## Temperature

model.jags.m3 <-  function()
{
  for (i in 1:n)
  {
    x[i] ~ dnorm(mu,tau)
  }
  variance  <- 1/tau
  desvpad   <- sqrt(variance)

  mu ~ dnorm(22,0.001)
  tau~ dunif(0,1)
}

## Humidity

model.jags.m4 <-  function()
{
  for(i in 1:n)
    {
    phi[i]  <--(L[i])
    zeros[i]~dpois(phi[i])

    a[i]    <-  x[i]/beta
    L[i]    <-  log(alpha) - alpha*log(beta) +
               (alpha-1)*log(x[i]) - a[i]^alpha
  }
  alpha ~ dunif(0.1,100)
  beta ~ dunif(0.1,1000)
}


## Weibull Mixture Model - With covariates ##

## Precipitation

model.jags.weib1 <-  function()
{
  c    <-  0.001
  for(i in 1:n)
  {
    phi[i]     <--(L[i])
```

```
    zeros[i]~dpois(phi[i])

    beta[i]     <-  gamma0 + gamma1 * (cov1[i]-2006) +
                    gamma2 * cov2[i] + gamma3 * cov2[i]^2
    logit(p[i]) <-  psi0 + psi1 * (cov1[i]-2006) + psi2 * cov2[i] +
                    psi3 * cov2[i]^2

    pdf[i]      <- log(alpha) - alpha * log(beta[i]) +
                   (alpha-1) * log(x[i]) - (x[i]/beta[i])^alpha
    surv[i]     <-  -(c/beta[i])^alpha

    L[i]     <-  (1 - delta[i]) * log(p[i]) +
                   + (delta[i]) * (log(1-p[i]) + pdf[i]- surv[i])
     }

  ## Priors
  alpha~dunif(0.1,2)
  gamma0~dnorm(0,1)
  gamma1~dnorm(0,10)
  gamma2~dnorm(0,10)
  gamma3~dnorm(0,10)
  psi0~dnorm(0,1)
  psi1~dnorm(0,10)
  psi2~dnorm(0,10)
  psi3~dnorm(0,10)

}

# Parameters JAGS
cov1  <-  dados$years
cov2  <-  dados$months
```

Appendix 5 - R Codes

```
###################
## No covariates
###################

## Tobit-Weibull Model

model.jags.weib11  <-  function()
{
c    <-  0.01
for(i in 1:n)
{
phi[i] <--(L[i])
zeros[i]~dpois(phi[i])
pdf[i]  <-  log(alpha)+log(gamma)+(gamma-1)*log(x[i])-(alpha*x[i]^gamma)
surv[i]  <-  -alpha*c^gamma

L[i]     <-  (1 - delta[i]) * log(p) + (delta[i]) * (log(1-p) + pdf[i]- surv[i])

cdfwei[i]     <-  log(1-p) + pdf[i]- surv[i]
res[i]        <-  -(cdfwei[i])
}
media   <- mean(L[])

## Priors
alpha~dunif(0,1)
gamma~dunif(0,2)
p~dunif(0,1)
}

## Tobit Exponentiated Weibull Model

model.jags.weib12  <-  function()
{
c     <-  0.01
for(i in 1:n)
{
     phi[i]  <--(L[i])
zeros[i]~dpois(phi[i])

a2[i]    <- -alpha*x[i]^gamma
a3[i]    <- 1-exp(a2[i])
```

```
pdf[i]   <-  log(alpha)+log(beta)+log(gamma)+(gamma-1)*log(x[i])+
          + a2[i]-log(a3[i]^(1-beta))
surv[i]  <-  log(1-(1-exp(-alpha*c^gamma))^beta)

L[i]   <-  (1 - delta[i]) * log(p) + (delta[i]) * (log(1-p) +
              pdf[i] - surv[i])

cdfwei[i]     <-  log(1-p) + pdf[i]- surv[i]
res[i]        <-  -(cdfwei[i])
}
media   <- mean(L[])

## Priors
alpha~dunif(0,1)
gamma~dunif(0,2)
beta~dunif(0,100)
p~dunif(0,1)
}

##  Tobit Generalized modified Weibull Model

model.jags.weib13  <-  function()
{
c     <-  0.01
for(i in 1:n)
{
phi[i]  <--(L[i])
zeros[i]~dpois(phi[i])
a2[i]   <- -alpha*x[i]^gamma
a4[i]   <- lambda*x[i]
a5[i]   <- a2[i]*exp(a4[i])
a6[i]   <- 1-exp(a5[i])

pdf[i]   <-  log(alpha) + log(beta) + (gamma-1)*log(x[i]) +
log(gamma+a4[i]) + a4[i] + a5[i] - log(a6[i]^(1-beta))
surv[i]  <-  log(1-(1-exp(-alpha*c^gamma*exp(lambda*c)))^beta)

L[i]   <-  (1 - delta[i]) * log(p) + (delta[i]) * (log(1-p) + pdf[i]- surv[i])

cdfwei[i] <-  log(1-p) + pdf[i]- surv[i]
res[i]    <-  -(cdfwei[i])
}
media <- mean(L[])
```

```
## Priors
alpha~dgamma(1,2)
beta~dgamma(3,2)
gamma~dgamma(1,2)
lambda~dgamma(0,0.1)
p~dunif(0,1)
}

###################
## With covariates
###################

## Tobit-Weibull Model

model.jags.weib21  <-  function()
{
c    <-  0.01
    for(i in 1:n)
    {
    phi[i]   <--(L[i])
    zeros[i]~dpois(phi[i])

    pdf[i]      <-  log(alpha) + log(gamma) + (gamma-1)*log(x[i]) -
                (alpha*x[i]^gamma)
    surv[i]     <-  -(alpha*c)^gamma

    logit(p[i]) <- psi0 +  psi1 * cov1[i] + psi2 * cov2[i]  +
                      psi3 * cov2[i]^2

    L[i]    <-  (1 - delta[i]) * log(p[i]) + (delta[i]) * (log(1-p[i]) +
                  pdf[i]- surv[i])

    cdfwei[i]  <-  log(1-p[i]) + pdf[i]- surv[i]
    res[i]     <-  -cdfwei[i]
    }
    media      <- mean(L[])

    # Priors
    alpha~dunif(0,10)
    gamma~dunif(0,2)
    psi0~dnorm(0,1)
    psi1~dnorm(0,100)
    psi2~dnorm(0,100)
    psi3~dnorm(0,100)
    }
```

```
## Tobit Exponentiated Weibull Model

model.jags.weib22  <-  function()
{
c     <-  0.01
        for(i in 1:n)
        {
        phi[i]   <--(L[i])
        zeros[i]~dpois(phi[i])
        a2[i]      <- -alpha*x[i]^gamma
        a3[i]      <- 1-exp(a2[i])

        pdf[i]      <-  log(alpha) + log(beta) + log(gamma) +
        (gamma-1)*log(x[i]) + a2[i] - log(a3[i]^(1-beta))

        surv[i]     <-  log(1-(1-exp(-alpha*c^gamma))^beta)

        logit(p[i]) <- psi0 +  psi1 * cov1[i] + psi2 * cov2[i]  +
                psi3 * cov2[i]*cov2[i]

        L[i]      <-  (1 - delta[i]) * log(p[i]) +
                        (delta[i]) * (log(1-p[i]) + pdf[i]- surv[i])

        cdfwei[i]     <-  log(1-p[i]) + pdf[i]- surv[i]
        res[i]        <-  -cdfwei[i]
        }
media   <- mean(L[])

## Priors
alpha~dunif(0,100)
gamma~dunif(0,20000)
beta~dunif(0,2)
psi0~dnorm(0,1)
psi1~dnorm(0,100)
psi2~dnorm(0,100)
psi3~dnorm(0,100)
}

## Tobit Generalized modified Weibull Model

model.jags.weib23  <-  function()
{
c   <-  0.01
for(i in 1:n)
```

```
{
phi[i] <--L[i]
zeros[i]~dpois(phi[i])
b1[i]  <-  exp(lambda*c)
b2[i]  <-  c^gamma
b3[i]  <-  exp(-alpha*b2[i]*b1[i])
b4[i]  <-  1-(1-b3[i])^beta
a1[i]  <- exp(lambda*x[i])
a2[i]  <- (x[i])^gamma
a3[i]  <- exp(-alpha*a2[i]*a1[i])
a4[i]  <- log(alpha)+log(beta)-(gamma-1)*log(x[i])
a5[i]  <- log(gamma+lambda*x[i])
a6[i]  <- lambda*x[i]-(alpha*a2[i]*a1[i])
a7[i]  <- (1-beta)*log(1-a3[i])
a8[i]  <- a4[i] + a5[i] + a6[i] - a7[i]
logit(p[i])    <- psi0 +  psi1 * cov1[i] + psi2 * cov2[i]  + psi3 * cov2[i]^2

L[i]   <-  (1 - delta[i]) * log(p[i]) + delta[i]*(log(1-p[i])
             + a8[i] - log(b4[i]))

cdfwei[i]     <-   log(1-p[i]) + a8[i] - log(b4[i])
res[i]        <-  -(cdfwei[i])
}
media   <- mean(L[])

## Priors
alpha ~dunif(0,100)
beta  ~dunif(0,100000)
gamma ~dunif(0,2)
lambda~dunif(0,10)
psi0  ~dnorm(0,1)
psi1  ~dnorm(0,100)
psi2  ~dnorm(0,100)
psi3  ~dnorm(0,100)
}
```